11TH ANNUAL
LANGUAGE TESTING RESEARCH COLLOQUIUM

HOSTED BY THE
DEFENSE LANGUAGE INSTITUTE ENGLISH LANGUAGE CENTER
SAN ANTONIO, TEXAS

MARCH 4, 5, & 6, 1989

ORGANIZING COMMITTEE

JACK UPSHUR, ABSTRACTS
DONNA ILYIN, COMMUNICATIONS
PANSY JOHNSON, LOCAL COORDINATOR
11TH LANGUAGE TESTING RESEARCH COLLOQUIUM
SAN ANTONIO, TEXAS
MARCH 4-6, 1989

PROGRAM

Saturday, March 4th:

10:00-12:00 REGISTRATION: in the lower level of the Emily Morgan Hotel. (All colloquium events will take place in the Emily Morgan Hotel).

1:15 OPENING: Welcome by Colonel Don W. Box, Commandant, Defense Language Institute English Language Center.

PRESENTATIONS:

1:30-3:00
1. Caroline Clapham. ELTSREV: Towards the Validation of an international EAP test.

2. Patricia Dunkel. Designing a computer-based ESL listening comprehension proficiency test: Prototype development efforts.

3:30-5:45
3. Lyle Bachman, Fred Davidson, Brian Lynch and Katherine Ryan. Content analysis and statistical modeling of EFL proficiency tests.

4. Margaret Des Brisay. The problem of the middle ground: Where do you draw the line?


Sunday, March 5th:

9:00-10:30
6. J.D. Brown. The place of ESL students in writing across the curriculum placement test battery.


11:00-12:30
8. Achara Wangsotorn. Domain-referenced tests for secondary and tertiary levels of education.

2:00-4:30 Minisymposium on language aptitude testing


Monday, March 6th:

9:00-10:30 15. Neil Anderson. Taking reading comprehension tests: What are second language readers doing?


2:00-5:00 19. Kyle Perkins and Charles Parrish. The determination of hierarchies among TOEFL reading comprehension items.


22. Grant Henning. Effects of short term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance.

Special thanks to the local committee: Susan Ferns, Al Boutin, Judy Goslin and Peggy Goitia, and to John P. Devine, OUIELC Chief of Curriculum Development.

Thanks also to Lyle Bachman, Donna Tlyin and Charles Stansfield for their invaluable planning assistance.
PRESENTATION 1: ELTSREV: TOWARDS THE VALIDATION OF AN INTERNATIONAL EAP TEST

PRESENTERS: Caroline Clapham and Gill Westaway

One of the most significant contributions to the testing of English for Academic Purposes (EAP) in recent years has been the English Language Testing Service (ELTS) test, jointly produced and administered by the British Council and the University of Cambridge Local Examinations Syndicate. Introduced in 1980, this test assesses the English language proficiency of overseas students wishing to study in the medium of English. Based on an analysis of what such students are perceived to do with language during their stay overseas, it measures candidates' general language ability and the skills needed for effective study or training. It includes modular subtests in different subject areas, and tests writing and speaking abilities directly. Test scores are reported according to bands of ability associated with performance descriptors.

From the ELTS validation studies it became clear that while the test was generally well-liked by its users, some aspects caused problems, particularly in the areas of administration and reliability. The test is therefore being revised, and in the process is becoming more international with the participation of Australian and Canadian test developers.

This paper reports on the progress of the Revision Project. It describes the development of the draft specifications and items, the procedures adopted for content validation, the ensuing revision of the specifications and test items, and finally the test piloting. It concludes by discussing the implications of the results of the ELTSREV research for EAP testing in general.
PRESENTATION 3: CONTENT ANALYSIS AND STATISTICAL MODELING OF EFL PROFICIENCY TESTS

PRESENTERS: Lyle F. Bachman, Fred Davidson, Brian Lynch, and Katherine Ryan

The Test of English as a Foreign Language (TOEFL) and the Certificate of Proficiency in English (CPE) are both used for the purpose of making admissions decisions about international students applying to educational institutions in North America and the United Kingdom. Despite this similarity of use, they represent two different approaches to language test development—the TOEFL exemplifying emphasis on psychometric considerations and the CPE representing focus on applied linguistic considerations. This study is aimed at investigating the comparability of these two tests.

The comparability of the two tests' content was examined through the analysis of each test's design facets—abilities and test methods—using frameworks of abilities and test methods proposed by Bachman (in press). The analyses of these facets also provided a basis for formulating hypotheses about how test takers might be expected to perform on different tests. In this study approximately 2,000 subjects in nine countries took the TOEFL, the SPEAK (a test of English writing similar to the Test of Written English), and either the CPE or a similar, lower-level test, the First Certificate in English (FCE). The statistical analyses of examinee responses were used to empirically test hypotheses about patterns of similar and dissimilar performance based on the a priori content analysis. In addition, examinee performance on these test batteries was used to investigate the comparability of test scores across the two tests.

This paper presents results of the content analyses of the reading sections of the TOEFL and the CPE or FCE, along with empirical results of analyses of examinee performance on these tests. The paper concludes with a discussion of the relevance of both the a priori analysis of test content facets and the statistical analysis of examinee responses for the development and use of language tests.
PRESENTATION 4: TWO TESTS ARE BETTER THAN ONE?

PRESENTER: Margaret Des Brisay

This paper synthesizes research conducted at three ESL pre-departure programs, one in China and two in Indonesia, funded by both Canadian and American development agencies.

Common to all these programs is the need for initial selection/placement procedures which predict who will be successful in their language training and the amount of training required. In the Indonesian programs, success is narrowly defined as achieving a score of 550 on an official TOEFL. In the China program, exit criteria are related to performance on the Canadian Test of English for Scholars and Trainees (CanTEST). The paper will focus on a study conducted at the Canada Indonesian Language Program (CILP) during an 18 week pre-departure course and briefly compare these findings to those from the other two programs.

In Indonesia, estimates of average gain scores on the TOEFL are commonly used as the basis for program planning. It is not surprising that problems arise given that, at the CILP, the standard deviation for the 41.5 point mean gain (between two official TOEFL's with 20 weeks of intervening language training) was 28 and gains in individual cases ranged from +113 to -10.

For a variety of reasons, administrators at all three programs wished to investigate the possibility that multiple measures might provide a better basis for decision-making. Data have been collected enabling comparisons to be made of examinees' performance on a variety of measures including in-house placement tests, oral interviews, writing tests, institutional and official TOEFL's and three versions of the CanTEST.

Overall, the data support the case for multiple measures. Moreover, they suggest many interesting areas of further research regarding: the greater predictive power of reading scores; the superior performance of newly-placed students over those who have "come up through the ranks" in continuing programs; effects of "coaching" on test scores.

Finally, the study supports the claim that any reasonably relevant test will serve to identify the truly proficient and those linguistically at risk. But, while the overall ranking of students may be similar on a given pair of tests, it appears that the closer a student's score is to the cut-off point, the more likely it is that a different test will result in a different decision being made about his future.

The problem of "where to draw the line" is not seen by the researchers as a statistical one. The fact is that in "the middle ground" there may not be enough difference in language proficiency to justify making different decisions about a candidates's future.

Test developers can either recommend higher cut-offs to safeguard the credibility of their tests (or permit admissions officials to do this for them) or set cut-offs low enough to give the benefit of the doubt to marginal cases.
Findings of these studies, then, form the basis for a policy debate as much as for further empirical enquiry.
PRESENTATION 5: THE USE OF INDIRECT MEASURES TO TEST L1/L2 WRITING ABILITY: SOME ISSUES

PRESENTERS: Doreen Ready and Robert Courschene

All engineering students, both L1 and L2, at the University of Ottawa who declare English their language of study must sit for an English test to determine if they must take the Second Language Institute's Engineering writing course. In the past, exemption was determined on the basis of a pass/fail mark on the Institute's writing test (a 300-400 word report on a specific subject marked for grammar, content, register and organization/style). As a result of a new proposal, the same group of students will now be evaluated using an indirect measure similar to the Graduate Record Exam. On the basis of their marks, students will be placed in one of four possible courses taught by the English Department: 1) grammar and composition, 2) workshop in essay writing, 3) literature and composition (prose or poetry option).

To determine the effect that this might have on the ESL students in the engineering population, both tests were administered to 127 second year engineering students. Comparisons were made regarding the placement decisions that would result in using each test.

The Pearson correlation between the scores on each test was .43 indicating that students were not ranked in the same way by the two tests. Cross tabulations were performed to see how many students would be successfully placed in the appropriate level English course based on the cut-off score of the direct measure test and on the various cut-off scores of the indirect test.

The success rate was judged only in terms of whether those students who received non-exempt status on the direct measure test were placed in either of the two lower level English courses (grammar and composition or the workshop in essay writing).

Four sub-groups were identified to see whether decisions were consistent regardless of whether English was the student's first or second language. The groups were native speakers of English (N = 31), non-native speakers who were francophones (N = 23), non-native speakers who were Chinese (N = 43) and other non-native speakers (N = 30).

The success rate for native speakers and francophones was 100%, for the Chinese sub-group, 74% and for other non-native speakers, 80%.

It appears that the indirect measure places native speakers and francophones appropriately and does less well with other ESL groups. The pattern occurs probably because with native speakers the assumption holds that in general, knowledge of grammar can be linked directly with writing skills. This appears also to be the case with the francophone group who in Canada are exposed to English in a manner that approximates native speaker experience. With other non-native speakers, some of the disparity between the level of grammatical knowledge and writing skills can probably be linked to the need to
demonstrate a minimum level of English competence through tests such as TOEFL where there is no direct measure of writing performance.

Further analyses will be done with the indirect measure test to see whether there are differences in response patterns between the groups.

As a result of the direct measure task two groups with specific writing problems were identified. First, a sub-group of native speakers who despite their acceptable scores on the indirect test had poor writing skills. Second, a group of non-native speakers whose writing was so poor that they needed an upgrading course before being placed in the credit-level program. An analysis of the writing of these two groups indicated that their errors were qualitatively different and would need different instructional strategies to be corrected. Without the use of the direct measure test, the existence of these special groups might not have been discovered until well into the writing program.
Three years ago, a writing across the curriculum program was implemented at the University of Hawaii at Manoa. As part of the program, policy decisions were made which mandated that placement testing for writing courses be coordinated for all undergraduate students on the campus - native speakers and ESL students alike. As a result, all incoming freshmen are assigned by placement tests to one of six composition courses: accelerated composition, regular composition, regular composition with required laboratory, remedial composition, regular ESL composition or preparatory ESL composition. The five hour Manoa Writing Test (MWT) requires students to write on two topics and revise both essays later in the day. Each student's work is then rated by at least two raters per topic for a minimum of four raters per student. This paper explores the place of ESL students in such writing across the curriculum placement testing.

The data are based on an entire year's administrations of the MWT to more than 1700 incoming freshmen, i.e., both native speakers of English and ESL students. In addition to the MWT, the ESL students were also required to sit the three-hour English Language Institute Placement Test (ELIPT), which has two subtests each for ESL listening, reading and writing skills. The results are described in terms of central tendency and dispersion for each of the groups of students on the MWT, the ELIPT and all sub-scores for each. As would be expected, the position of the ESL students is clearly low in the overall distribution of MWT scores. The relative reliability of the MWT and the ELIPT is examined to determine the appropriateness of each for making decisions about the placement of ESL students. In addition, the degree to which the ESL students form a separate population within the total student body is explored by using one-way ANOVA on the MWT scores and by examining the intercorrelations among all sub-scores of both tests. The results are discussed in terms of how ESL testing and decision making was affected by these sweeping university-wide policies, as well as in terms of how we went about defending our students against the problems that inevitably arose from ignorance of their special ESL needs.
PRESENTATION 7: THE CONSTRUCTION AND VALIDATION OF A PERFORMANCE-BASED TEST BATTERY

PRESENTER: Michael Milanovic

Most research work in test validation has taken as its starting point either language tests commonly in use, such as TOEFL or the FSI oral interview. Furthermore, rather strange language tests have sometimes been constructed to make up necessary components in validation procedures such as the multimethod-multitrait matrix. While this is understandable and even necessary from the research point of view, it is not very desireable from the educational standpoint.

The battery of tests discussed in this paper was constructed within the context of a large language teaching institute in Hong Kong. While it was seen as absolutely necessary that the tests could be subjected to rigorous validation procedures, it was also vital that the tests fitted into the pedagogic environment of the institute, since they were intended to have a positive washback effect on teaching.

It was demonstrated through classical test statistics that performance-based tests, when constructed within the context of a teaching institute, could perform as well as or better than any multiple-choice test might. In addition exploratory factor analysis suggested that the tests were indeed testing different skills, as it was hypothesized they should.

Further factor analyses of the items in the Listening Component of the tests identified subskills which closely matched the intuitively hypothesized subskills upon which the items had been based.

The paper suggests that more emphasis in test validation in the future should be placed on the nature of test items used, and that researchers should be less READY to employ tests which are in themselves of questionable validity.
PRESENTATION 8: DOMAIN-REFERENCED TESTS FOR SECONDARY AND TERTIARY LEVELS OF EDUCATION

PRESENTER: Achara Wangsotorn

Objectives The research had the following objectives:

(1) To develop unitary and integrative-skills standardized domain-referenced test items from the determined domains of English use via the sound modality and the graphic modality,
(2) To study, analyze and determine the factors of communicative competence in using English of Thai students in the lower secondary, upper secondary and tertiary levels of education, and
(3) To determine the common levels of English proficiency of Thai students.

Procedures The following steps were used in the research:

(1) Analyzed the objectives, syllabuses and tests to set the domain-referenced test specifications.
(2) Developed and analyzed the test items by a domain-referenced model of item analysis, improved to meet the standards and administered to the sample groups: 1,222 secondary students obtained by the multi-stage stratified random sampling method and 493 university students obtained by the stratified random sampling method.
(3) Parametric statistics were used to analyze the student levels of English proficiency which were set at 5 levels from very weak requiring extensive remedial work to good - very good exempted from remedial work.
(4) Pearson-Product-Moment Correlations and Exploratory Principal Component Analysis with Varimax Rotation were used to study about the factors of communicative competence.

Findings Research findings revealed that the students had better proficiency in sound modality and language components than in the graphic modality. They were especially weak in writing and in integrated skills of reading combined with writing. The study about the traits of communicative competence indicated that grammar, vocabulary and phonology shared common variances with the sound and graphic modalities indicating that they could be subsumed under the language modalities. It was also discovered that the number and traits of language learning factors differed among the levels with the lower secondary level having more sound modality factors while the tertiary level had more graphic modality factors. SSQ analysis revealed that sound modality shared greater variances with domain-referenced proficiency at the ratio of 1.8:1, and that both modalities had common variance with domain-referenced proficiency by approximately 50%.
PRESENTATION 9: ASSESSMENT IMPLICATIONS OF A CONTENT-BASED CURRICULUM: THE ROLE OF SELF-ASSESSMENT

PRESENTERS: Kathleen M. Bailey, Peter A. Shaw, and David Tsugawa

In this presentation we will examine the role of self-assessment procedures in a curriculum development project, which was funded by a three-year grant from the Pew Memorial Trust. College students enrolled in "typical" language courses were compared to those enrolled in innovative "content courses" on economics and political science. These courses were taught in various target languages (German, Spanish, French or Chinese). Students in both the control and the experimental groups were pre-tested with the ACTFL Oral Interview procedure, which thus serves as one criterion measure against which the students' pre-course self-assessments are compared. Other pre-course measures included cloze passages and target language writing samples.

The self-assessment questionnaire was based on two domains: (1) general target language use, and (2) academic target language use. Items about general target language use were designed specifically to tap those communicative skills and functions assessed in the ACTFL Oral Interview. In the wide-scale curriculum development project, the students' academic self-assessments will be interpreted in terms of data from videotapes and ethnographic observations of the students' classroom behavior, as well as entries from the students' language learning journals. In this presentation, however, we will focus on the relationship between the students' self-ratings and the various pre-course measures: (1) the interviewers' independent ACTFL oral proficiency designations of the students, (2) ratings of the students' target language writing samples, and (3) their performance on the cloze passages. Correlations among the various criterion measures will also be examined. Implications for self-assessment (e.g., its criterion validity, its practicality, and its possible benefits to the language learners) will be discussed in the light of previous research on this topic, including papers presented at the 1988 Language Testing Research Colloquium.
PRESENTATION 10: A RATIONALE FOR A REEXAMINATION OF LANGUAGE APTITUDE TESTING

PRESENTER: Charles W. Stansfield

For most practitioners in second language education, the measurement of language aptitude is a non-issue. First, recent research in second language acquisition focuses on how one learns/acquires a second language, regardless of any "aptitude" for it. Moreover, interest in second language education tends to concentrate on methodology for successfully teaching a foreign language to anyone, again regardless of "aptitude." Thus, we now have a plethora of pedagogical approaches such as TPR, the Silent Way, the Natural Approach, and so on. Finally, in an era of declining foreign language student enrollments, encouraging the concept of language aptitude (i.e., either one can learn a foreign language or one cannot, and most Americans feel they cannot) seems detrimental to the field at best.

Nevertheless, accurate measuring of aptitude for learning modern foreign languages remains a definite concern to many U.S. government agencies, where aptitude testing is commonly used in selection for and/or placement in language training programs. Because of the neglect of this area in the larger field, however, very little has been done to improve the accuracy of language aptitude testing since the appearance of Carroll's Modern Language Aptitude Test (MLAT) over two decades ago. This is despite advances in knowledge from second language acquisition research on how one learns a second language and work in the cognitive realm of what is involved in learning a second language.

Thus, in the Spring of 1987, the Interagency Language Roundtable (ILR) of the U.S. Government discussed the need to pursue a major initiative in the testing of foreign language learning aptitude. From that initial discussion flowed the plans for a conference on language aptitude research and testing. In September, 1988, the ILR Invitational Symposium on Language Aptitude Testing was held at the Foreign Service Institute Language School in Rosslyn, Virginia.

The symposium had three goals. The first was to ensure that the ILR discussion of language aptitude testing would be infused with the latest research on relevant topics. The second was to provide a forum for cooperation among the governmental agencies in the area of language aptitude measurement to the fullest extent possible. The third and perhaps most important was to work on the production of a written set of recommendations for future work in the area of language aptitude testing.

This minisymposium on language aptitude testing at the 11th Language Testing Research Colloquium will share with the broader field some of the insights gained from the ILR symposium and open up the discussion of new initiatives in language aptitude testing to a broader audience.
PRESENTATION 11: PRELIMINARY INVESTIGATION OF THE RELATIONSHIP BETWEEN VORD, MLAT, AND LANGUAGE PROFICIENCY.

PRESENTERS: Thomas S. Parry and James Child

The purpose of this paper is to report preliminary findings of a joint exploratory study conducted in 1987-88 between the Department of Defense and the Central Intelligence Agency to shed further light on the psychometric properties of a new artificial language aptitude test known as VORD. Based on Turkic language structural typologies, VORD was developed in response to a need for an instrument better predictive of success in learning non-Indo-European languages. Such commonly used language aptitude tests as the Army Language Aptitude Test (ALAT) and the Modern Language Aptitude Test (MLAT) were developed and validated in the late 1950's to predict learner success particularly in Western European languages. The present study examines whether or not intercorrelations exist between VORD and MLAT subtests and how much reliability each measure has in predicting end-of-training language proficiency outcomes.

The study was carried out in two phases. In phase one, the following research question was posed: Do significant correlations exist between MLAT and VORD subtests? From this question, two tangentially related questions were posed: 1) Is there a significant correlation between performance on MLAT/VORD and learners' perceived aptitude to learn foreign languages?, and 2) Is there a relationship between aptitude test performance and such learner variables as time-in-training, age, gender, level of motivation, and overall satisfaction with language training? For phase two of the investigation, the following questions were posed: 1) Do significant correlations exist between learner performance on MLAT/VORD and outcomes on end-of training oral and reading proficiency tests?, and 2) Which subtests of the MLAT and VORD, either individually or in combination, are the strongest predictors of oral and reading proficiency test outcomes?

Thirty-six subjects (17 male and 19 female), enrolled in a government language program, volunteered to participate in the study. All were native speakers of English ranging in age from 21 through 56 years. Many had completed several years of service abroad and were learning their second and in some cases their third language. Subjects completed the VORD, MLAT, a questionnaire, and end-of-training oral and reading proficiency tests. Data were analyzed using the standard correlation and regression programs of the Statistical Analysis System.

For phase one of the study, data analysis revealed that significant moderate correlations exist between MLAT and VORD composite scores (r=0.695, p<.01) and that correlations between MLAT and VORD subtests ranged from low to moderate (r=0.2 to 0.68) with the moderate correlations being significant (p<.01). On the question of learners' perceived aptitude, subjects viewed themselves as average to slightly above-average language learners on a range from poor to superior. They also tended to score in the average range on the MLAT (per government norms). The correlation between these variables was found to be moderate and significant (r=0.727, p<.001). Subjects scored in the 43rd percentile on the VORD (no norms) resulting in a mild but significant correlation with learner perceived aptitude (r=0.450, p<.05). No significant
correlations were found to exist between the variables of age, level of motivation, overall satisfaction with language training, and MLAT/VORD subtest and composite scores. The time-in-training variable did not correlate with any of the VORD subtests, but was found to be significantly correlated with MLAT composite scores ($r=0.591$, $p<.05$) and MLAT subtest III ($r=0.621$, $p<.01$).

Data analysis for phase two revealed mild correlations between performance on the MLAT (composite scores) and speaking proficiency ($r=0.476$, $p<.01$) and reading proficiency ($r=0.446$, $p<.01$). Mild correlations were also detected between VORD composite scores and speaking proficiency ($r=0.463$, $p<.01$) and reading proficiency ($r=0.345$, $p<.05$). Combining VORD/MLAT subtests in a stepwise regression analysis, MLAT subtest II, phonetic script, was found to be the strongest predictor of reading proficiency while MLAT subtest III, spelling clues, was the strongest predictor of speaking proficiency. Of the four VORD subtests, the Sentences subtest proved to be the strongest predictor of both language skills. MLAT composite scores were significantly better overall predictors of both speaking and reading language proficiency than VORD composite scores.

Although limited by small sample size, the present study provides evidence that MLAT is more effective than VORD as a predictor of language proficiency outcomes. There is evidence, however, that VORD may be a better predictor of learner outcomes in carrying out such discrete language tasks as grammatical analysis. This leads the researchers to conclude that language aptitude is more than a unidimensional construct.
PRESENTATION 12: PREDICTORS OF SUCCESS IN AN INTENSIVE FOREIGN LANGUAGE LEARNING CONTEXT: PREDICTING LANGUAGE LEARNING AT THE DEFENSE LANGUAGE INSTITUTE FOREIGN LANGUAGE CENTER

PRESENTER: JOHN A. LETT, JR.

For many years it has been hypothesized that aptitude for learning foreign languages both exists and can be measured. Indeed, meaningful relationships have long since been established between language training outcomes and scores on various tests which purport to measure language learning aptitude. However, as more than two decades of research have shown, cognitive ability, even if defined and measured with reference to specific learning domains, is by no means the only learner characteristic which can be meaningfully linked to learning outcomes. Accordingly, recent research conducted within the military language training context, like that conducted outside the government arena, has addressed a broad array of both cognitive and non-cognitive individual characteristics. The purpose of this paper is to describe the preliminary outcomes of one such study being conducted jointly by the Defense Language Institute Foreign Language Center (DLIFLC) and the US Army Research Institute (ARI).

This study, known as the Language Skill Change Project (LSCP), is a longitudinal study whose principal objective is the tracking of language skill change over time; however, it was designed to support an investigation of variables associated with initial foreign language acquisition as well. Predictor data were gathered from 1903 selected Army enlisted personnel beginning a DLIFLC basic course in Korean, Russian, German, or Spanish, between February, 1986, and August, 1987. Criterion measures were scores on the Defense Foreign Language Proficiency Test, III (DLPT III), plus a dichotomous variable indicating academic attrition or completion of the course. After various data reduction analyses, data were submitted to multiple regression analyses using a forward progression, forced order of entry approach, with variable blocks being entered in an order of increasing implementation costs.

Results of the regression analyses indicate that proficiency attainment can be predicted better than academic attrition, that speaking scores are less predictable than listening and reading scores, and that the relative importance of given predictor variable sets varies across languages and skill modalities. In general, the role of cognitive ability was supported, with general ability being more valuable in predicting success in the less difficult languages (Spanish and German) and language learning aptitude being more useful in the case of the more difficult languages (Russian and Korean). Non-cognitive measures also contributed significant amounts of variance: learning strategies were consistently useful in the prediction of listening and reading skills; student attitudes and motivation measured both prior to and during language training were related to attained listening scores; the measures of attitudes and motivation collected during training (but not the pre-training measures) were related to reading and speaking scores; and prior foreign language training was significantly related to listening scores in the less difficult languages, and to speaking scores in German only. These
preliminary findings support the importance of both cognitive and non-cognitive variables, but also suggest that current explanatory models of classroom language may not be fully operative in all language contexts.
PRESENTATION 15: TAKING READING COMPREHENSION TESTS: WHAT ARE SECOND LANGUAGE READERS DOING?

PRESENTER: Neil Anderson

The purpose of this presentation is to evaluate the strategies that second language (L2) students use while taking a standardized reading comprehension test. This presentation is part of a larger research study that compared the reading strategies of L2 readers in two contexts: first, the strategies used to take a standardized reading comprehension test and second, the strategies used during academic reading. This presentation will only present the results of the strategies used during the testing context.

The research was designed to answer the following questions:

1. What reading comprehension strategies do second language readers use while taking standardized reading comprehension tests?
2. How do the use of strategies differ according to the level of L2 proficiency?

The participants for this research were twenty-eight native Spanish speaking students enrolled at the Texas Intensive English Program in Austin, Texas, during Fall semester 1988.

All subjects took forms A and B of the Descriptive Test of Language Skills -- Reading Comprehension Test, a standardized reading-comprehension test consisting of fifteen reading passages each followed by two to four multiple-choice (four choice) comprehension questions for a total of forty-five questions. There are two forms of the test, Forms A and B. Both forms of the test were used in the research. The Descriptive Test of Language Skills -- Reading Comprehension Test is used at many colleges in the United States as a screening and placement test for native and non-native English speaking students in remedial and ESL reading programs (Block, 1986; Segel, 1986). The test was specifically designed to differentiate among lower-proficiency students (Educational Testing Service, 1985, p. 4).

The administration of the first form of the test was given under normal testing conditions. The second form was modified somewhat allowing the participants to give think-aloud protocols to describe their reading and testing strategies. The think-aloud protocols were given in English, Spanish or a combination of the two languages. The administration of the two forms of the standardized test was separated by the administration of the Textbook Reading Profile (TRP). The TRP was used to gather information about the students' reading strategies during academic reading tasks. This information will not be presented.

The data were analyzed in the following manner. The first form of the DTLS provided a reading comprehension score on which all participants can be divided into two groups: high comprehenders and low comprehenders. The strategies of these two groups were compared to determine the type of strategies each group invokes during the administration of the second form of
the STLS. The data were also analyzed by dividing the participants into three levels of language proficiency (beginning, intermediate, and advanced) according to the results of the placement exam for the Texas Intensive English Program. The strategies of these three groups were compared to determine which strategies the readers invoke at each of these levels of language proficiency.

This research provides insight into the strategies that underlie success of L2 readers taking reading comprehension tests by providing information on how students actually take reading comprehension tests as opposed to what they think they are expected to be doing.
PRESENTATION 16: THE TAKING AND RATING OF SUMMARY TASKS

PRESENTER: Andrew D. Cohen

Curiosity concerning what might make summary tasks unreliable, plus faith in the value of verbal report in providing data on cognitive strategies, motivated the undertaking of a small-scale study which had as its purpose to investigate: (1) the ways in which respondents carry out summarizing tasks on a reading comprehension test and (2) how raters deal with the responses. The research proposal for this study was reported on at the 9th LTRC in Florida (1987).

The respondents were five native Portuguese speakers who had all completed an EAP course with an emphasis on reading strategies, including summarizing. Two EAP course instructors who typically rated the EAP exams of summarizing skill also participated in the study as raters. The respondents had to complete a model EFL proficiency test, included in an exemplary testing package for Brazilian EFL teachers. The instrument included the writing of three summaries, two based on short texts and one based on a longer text. Respondents were also asked to indicate any processing problems they encountered. Respondents were requested to provide self-observational and self-revelational data during the taking of the test, and after the testing session they responded to a questionnaire dealing with their attitudes toward the EAP course, their use of English, and their opinions about this particular test. The raters also provided self-observational and self-revelational data while assessing the tests, and completed a questionnaire concerning their experiences marking the sample test.

The respondents had little difficulty identifying topical information, yet had difficulty distinguishing superordinate, non-redundant material from the rest, due in large part to an insufficient grasp of foreign-language vocabulary. In addition, they did not have a good sense of balance with respect to how much to delete. Either they were too vague and general or too detailed. While there was some concern for coherence production, there appeared to be relatively little attention paid to producing thoroughly coherent and polished summaries. Whereas successful summarizing requires the effective use of both reading and writing strategies, this study would suggest that respondents may exercise more strategies in reading the source texts than in writing text summaries.
PRESENTATION 17: THE UNDERLYING FACTOR STRUCTURE OF L2 CLOZE TEST PERFORMANCE IN FRANCOPHONE, UNIVERSITY-LEVEL STUDENTS: CAUSAL MODELING AS AN APPROACH TO CONSTRUCT VALIDATION

PRESENTER: Carolyn E. Turner

In the field of language testing, the cloze procedure has been widely adopted as a measure of overall second language (L2) proficiency. Despite the extensive research carried out on the cloze, there continues to exist a theoretical problem: that of construct validity. This paper describes a study which begins to address this problem.

The study investigates the underlying factor structure of L2 cloze test performance through the use of causal modeling as an approach to construct validation. The research question under investigation is: How is performance of L2 cloze tests dependent upon the following hypothetical constructs (causal variables) that are assumed to reflect knowledge pertinent to successful cloze performance: cloze-taking ability, knowledge of language, content domain, and knowledge of contextual constraints?

Eight cloze tests reflecting the posited factors were constructed and administered to 182 Francophone, university-level students. The factors were examined separately and in combination through a model building process which included model fitting and model comparison. The purpose was to confirm a theoretical model which best explained the intercorrelations among the cloze tests (i.e., the extent to which the cloze response patterns corresponded to the hypothesized predictions). The intent was to identify distinct factors contributing to L2 cloze test performance.

The results of the study confirm a model composed of three orthogonal factors as the best explanation of the data. The results indicate that cloze performance is dependent upon language factors (a second language factor or a first language factor) and nonlinguistic-specific knowledge related to cloze-taking ability that crosses over linguistic boundaries. The model building process also confirmed that the addition of the remaining posited factors did make a difference in accounting for cloze performance (i.e., did improve the relative fit of the model), but not a significant difference.

Cloze has been considered as an overall L2 proficiency measure. This study empirically demonstrates that factors other than language are significantly contributing to cloze test performance. Those involved in interpretation of cloze scores need to take this into account.

This study begins to address the question of cloze test construct validity. In doing so, however, more questions are raised. Even though a model containing two language factors and a factor of nonlinguistic-specific knowledge was confirmed as the best explanation of the data, it appears the relationship between these factors is not a static one in that it varies from one context to another (i.e., from one cloze test to another as well as from L2 cloze tests to L1 cloze tests). This leads one to believe that there is an interplay between the factors of language and the factor of nonlinguistic-specific knowledge brought out by different language test data. In order to address this query, it would seem appropriate to partition the
contributing factors into multiple contributing factors. Which multiple factors to include, however, is not immediately evident. The results of this study indicate that the division between linguistic and nonlinguistic-specific knowledge is not clearly defined.

In addition to addressing cloze test construct validity, this study demonstrates the potential of causal modeling as a theory-driven procedure.
PRESENTATION 18: A CONSTRUCT VALIDATION STUDY OF LISTENING AND READING COMPREHENSION

PRESENTER: Gary Buck

It is quite common, among both language teachers and language testers, to consider listening and reading as two separate language skills. However, the factorial structure of language proficiency is far from clear, although there seems to be a consensus on favor of some form of the divisible-competence hypothesis. The obvious question then is in what way is language proficiency divisible? In the case of listening and reading should we think in terms of one factor (or trait), namely comprehension, or two factors (or traits), listening comprehension and reading comprehension? This question is of great theoretical and practical importance not only to language testers, but also to teachers and materials writers, yet the empirical evidence is at best inconclusive.

The present study is designed to address this question by means of a multitrait-multimethod construct validation study. Two traits, listening comprehension and reading comprehension, and four methods, short answer questions, fill-in-the-blanks, translation and multiple choice questions, were included in the analysis. The eight tests constructed were administered to over 400 Japanese university students. The resulting correlation matrix is examined by means of Campbell and Fiske (1959) criteria, ANOVA and LISREL. Results indicate that listening comprehension and reading comprehension are two separate, but highly correlated, traits.

The paper also examines the actual tests used in the study in an attempt to isolate those variables which could have accounted for the differences between listening and reading comprehension, and these are discussed with special reference to the problems of testing listening comprehension. Finally, suggestions are made for further research.
PRESENTATION 19: THE DETERMINATION OF HIERARCHIES AMONG TOEFL READING COMPREHENSION ITEMS

PRESENTERS: Kyle Perkins and Charles Parish

Item responses from the reading comprehension subtest of the TOEFL administered to Southern Illinois University students (at the Center for English as a Second Language) will be submitted to the following measures to determine the hierarchies that exist among those items:

1) Bart and Krus’s (1973) ordering-theoretic method of analysis (to confirm or disconfirm two-item ordering); a matrix indicating the percentage of disconfirmatory response patterns for all of the two-item trees possible will be constructed;

2) Bart and Read’s (1984) statistical test will be used to determine whether the identified prerequisite relations are a matter of chance. (This procedure employs Fisher’s exact test and its unit-normal deviate approximation.) Our generalized descriptions of the prerequisite relations will be based on the constructs of (a) the relationship of each item to the structure of the test, (b) the reader’s prior knowledge, and (c) the nature of the cognitive processes required to answer the questions;

3) Meyer’s (1975) prose-analysis system will be used as a guideline for discussing how the items relate to the structure of the text. Meyer’s proposal of superordinate levels--overall organization derived from main ideas, middle level of supporting ideas, lowest level of minor details--provides the framework of rhetorical relationships (consisting of the 5 groups cause/effect, comparison, collection, description, and response) by which to judge student understanding of text;

4) Schlesinger and Weisner’s (1970) facet design, and Pearson and Johnson’s (1978) classification scheme and operational rules will be used to determine different comprehension categories to ascertain whether the pairs of items were dependent on student’s prior knowledge;

5) Carpenter and Just (1986) provide the framework for our consideration of cognitive processes. (The three issues germane to our "prerequisite relations" are decoding, processes of syntax and semantics, and specific linguistic structures.)
PRESENTATION 20: STRATEGIC COMPETENCE AND THE SPEAK TEST: AN EXPLORATION OF CONSTRUCT VALIDITY

PRESENTERS: Rosa Fagundes and Dan Douglas

This project is a study of the construct validity of the SPEAK Test with reference to one model of communicative language ability (viz. Bachman, In Press). Specifically, we are analyzing transcribed protocols of performance by non-native and native speakers of English on the SPEAK, and examining them for grammatical, textual, illocutionary and sociolinguistic exemplars of the planning process in responses to selected items on the test. We will look for associations between planning and execution strategies and performance as represented by test scores. We wish to discover whether some strategies may be favored over others, whether subjects from different disciplines may favor different strategies, and whether NNS subjects differ in any systematic way from NS subjects in planning and execution strategies.

SUBJECTS. The subjects are 60 graduate students in Chemistry, Math and Physics (20 in each field) at Iowa State University. Half of each group are NNS and half NS of English. The NNS subjects are Chinese and Korean and represent a range of SPEAK test performance from 130 to 260, with a mean of 190. The NNS subjects took Form 1 of the SPEAK test between 1984 and 1985 as a requirement for holding teaching assistantships in their departments; the NS subjects took Form 1 of the SPEAK voluntarily for this project during one week in October, 1988.

DATA. The data consist of transcribed protocols of subjects' responses to three items on the SPEAK: picture sequence, single picture and open-ended questions. The NNS subjects' TOEFL scores and biographical data such as sex, age, and field will also be used in the analysis.

ANALYSIS. We are analyzing the protocols for certain exemplars of the four competence areas mentioned above, to investigate the planning and execution process. As evidence of grammatical competence, we are looking at, for example, the use of definite/indefinite articles and passive/active verbs. As evidence of textual competence, we are looking at such features as

1) reference and 2) rhetorical strategies. We are studying the use of negation as one piece of evidence of illocutionary competence, since negation has been hypothesized to be related to direct/indirect performance of illocutionary acts (Leech 1983). Finally, we are looking at indicators of sociolinguistic competence such as the use of varieties of styles marked by slang or technical vocabulary and elision, assimilation and other simplification techniques. Related to this, we are also noting the degree of precision and detail (cf. the Gricean maxim of quantity) provided by subjects.

OUTCOME. We will compare NNS performance on terms of these and other features with that of our NS subjects, and also attempt to relate these elements of planning and execution to test score level, major field and language background. The outcome of the research we hope will be a better understanding of the construct validity of the SPEAK test, with reference to one model of communicative language ability.
In the Colloquium, we will present results of our analysis to date, as well as selected data for discussion, seeking the advice of colleagues on its analysis and interpretation.
PRESENTATION 21: PERSON DIMENSIONALITY, PERSON ABILITY AND ITEM DIFFICULTY IN WRITING

PRESENTERS: Liz Hamp-Lyons and Sheila Prochnow

It has long been known that individuals receive the same test scores for different reasons. Test reports which include sub-test scores provide some information about the components which make up an individual’s language proficiency, but even sub-test scores obscure differential performance and ability across individuals. Nor is it realistic to propose reporting item scores separately.

When a single score is reported, on a test or a component of a test, score consumers assume that this score "represents" performance on the test items; that is, that the score is somehow typical of the individual’s performance across all the items. On dichotomous items this is usually the case; the typical response pattern on a 50-item test would be success >> sporadic failure >> sporadic success >> failure (a ++++0+0+++00+0000 pattern): I call this a ‘marked’ pattern, the reported score does not pass on this information. The situation is even less simple on a test using a rating scale. First, there are fewer items -- often only a single item. Second, rating scales are typically used to measure complex language behaviors such as speaking ability and writing ability -- behaviors which it is generally acknowledged are composed of many features and skills. Holistic rating scales seek the ‘center of gravity’ in the rater’s response to the behavior to be assessed; analytic rating scales take the behavior apart and attempt to break it into its main components, permitting performance variation on different components to be revealed -- but typically they then have to put the components together again in order to produce a single score for bureaucratic use. In research in Britain I found that perhaps 2 in 10 testees exhibit ‘marked’ performance on component measures of a writing test; Ben Wright has had a similar experience in the U.S.A.

I will describe an innovation in the MELAB (Michigan English Language Assessment Battery) in which optional codes are added to a holistic rating to report ‘marked’ performance on the writing sub-test, while ESL writers with ‘flat’ performance receive a holistic rating without codes. I will examine data on the frequency of marked performance (person multidimensionality); the relationship between marked performance, holistic score and language proficiency level (person ability); and the relationship between marked performance and item difficulty. I will suggest that this system offers a manageable way of acknowledging that not all samples of writing can be defined by a single holistic score while also acknowledging that many can. I will report inter-rater and intra-rater reliabilities for holistic scores, for holistic scores on samples of essays with and without marked performance, and for identification of marked performance.
NOTE: The following two abstracts are of papers which will not be presented at the Colloquium.

VALIDITY AND RELIABILITY OF THE PORTUGUESE SPEAKING TEST

Charles W. Stansfield and Dorry M. Kenyon
Center for Applied Linguistics
1118 22nd. St. NW
Washington, DC 20037

The Portuguese Speaking Test (PST) was developed by staff at the Center for Applied Linguistics with support provided through a grant from the U.S. Department of Education. The PST was developed to provide a valid and reliable measure of Portuguese oral language proficiency in situations where it is not possible to provide a trained oral proficiency interviewer in Portuguese. Thus, the PST is a semidirect test that can be administered to groups of examinees through a test tape and a test booklet in a language laboratory setting. The examinee hears the instructions and test questions on the test tape, and answers each question orally on a separate tape. The test booklet contains visual stimuli in the form of pictures as well as text questions.

The PST is modeled on the American Council on the Teaching of Foreign Languages (ACTFL) and the Interagency Language Roundtable (ILR) Oral Proficiency Guidelines and the accompanying oral proficiency interview (OPI). It follows the interview format in that it begins with a warm-up, which is a simulated conversation between two people who meet each other for the first time. It then progresses to more difficult tasks that determine the examinee's ability to carry out various tasks and functions in Portuguese. These include giving a set of directions, describing a place, narration in present, past, and future time, presenting different points of view on a controversial topic, defending a particular point of view, arguing, persuading, and addressing a group of people in a formal setting. The entire test takes about 45 minutes to administer and produces about 25 minutes of speech on the part of the examinee. Upon completion of the test, the examinee's response tape is sent to a trained rater who assigns a proficiency rating on the ACTFL/ILR scale.

In order to determine the reliability and validity of the PST, a study was conducted during May 1988. In this study, 30 adult students of Portuguese, at a broad range of abilities and of widely diverse backgrounds, were administered two different forms of the PST as well as the OPI during a two-day period. An ACTFL certified and an ILR certified rater were used. For a given subject, the OPI was administered and the subject rated by one of these two raters. Each OPI was recorded and scored at a later date by the other rater. Each rater also scored all the PST tapes. Thus, each OPI and each form of the PST was scored independently by each rater using the ACTFL/ILR scale. The data collected from the study serves as the basis of the paper.
The results of the study indicate that the PST is highly reliable for a free response test. The three different forms of the test showed an average parallel form reliability of .99 when scored by one rater and .96 when scored by the other. The test can also be scored reliably by different raters as evidenced by the fact that interrater reliabilities for the different forms ranged from .93 to .98.

The results of the study also indicate that the PST is valid as a test of oral communication ability. Correlations between the PST and a live interview ranged from .90 to .96, depending on the PST form and the rater used to score the interview. The average correlation between all ratings on the PST and all ratings of the live interview (180 observations) was .93. This figure takes into account both same rater and different rater scoring. The magnitude of the validity coefficients supports the criterion-related and construct validity of the PST and its use as an alternative to the OPI in situations where a live interview is not possible.

Subjects also completed a questionnaire designed to assess and compare attitudes toward the two measures. While examinees provided positive ratings of both tests, they preferred the human contact involved in the live interview.
MEASURING COMPONENTS OF LINGUISTIC COMPETENCE

Peter Skehan

Current models of communicative competence (e.g. Bachman forthcoming) are wide-ranging in scope, yet they generally contain a linguistic/syntactic component. Even so there are many gaps in our knowledge of how such a syntactic competence is structured and how it can be tested. This paper examines ten aspects of syntactic competence and examines how effective each of them is in discriminating between learners at different proficiency levels. The ten areas will be examined through traditional reliability analysis, item-response theory, and correlational/factor analysis. It will be argued that test scales based on transitivity as a syntactic area operate more successfully (i) at different proficiency levels, and (ii) with less influence from subjects’ L1 and previous learning conditions.

In addition three areas were used in a multitrait-multimethod design - the article system, transitivity and sentence structure, and logical connectors. Each was assessed through multiple-choice and cloze formats. The paper will report on the respective contributions of trait and method factors in each case.

The results will be discussed in terms of the feasibility of using different areas of syntax as the basis for proficiency or diagnostic testing. The different areas will also be considered for suitability for inclusion in an index of general second language development.

Bachman L. (forthcoming), Fundamental Considerations in Language Testing, Addison Wesley
Re: LTRC 11, March 1989,

Facts - Fk

Items

What is saliency and what does it imply? How do we assume that grade familiarity? For the less familiar facts.

For each item:
- Determine most salient single fact
- Determine 2nd most salient fact
- Determine 3rd in a fact

(Where saliency is defined as average feeder rating closest to either pole.)

Form 3 sets of item parcels; use as control groups.