LANGUAGE ASSESSMENT IN EDUCATION

Thirteenth Annual
Language Testing Research
Colloquium

March 21-23, 1991
Educational Testing Service
Henry Chauncey Conference Center
Princeton, New Jersey
SPECIAL ACKNOWLEDGMENTS

COLLOQUIUM ORGANIZERS

Elana Shohamy and Grant Henning

ABSTRACT READERS

Lyle Bachman
Kathleen Bailey
Frances Butler
Eduardo Cascallo
Andrew Cohen
Fred Davidson
John H.A.L. de Jong
Dan Douglas
Liz Hamp-Lyons
Grant Henning
Thom Hudson
John Oller, Jr.
Don Porter
Elana Shohamy
Charles Stansfield
Diane Wall
Marjorie Wesche

PROGRAM TYPING AND DISTRIBUTION

Joanne Farr

MAILING LIST PREPARATION AND MAINTENANCE

Joanne Farr

CONFERENCE CENTER REGISTRATION AND MATERIALS

Cheryl Campbell
Joanne Farr
Beth Greczyn
Donna Natriello
Debbie Newton
Elaine Rick
1991 LANGUAGE TESTING RESEARCH COLLOQUIUM
THURSDAY, MARCH 21
CHAIR: Grant Henning

8:00 - 9:00 Registration
9:00 - 9:15 Opening Remarks - Henry Braun, Vice President for Research Management, ETS
9:15 - 10:00 Plenary
   Toward a Test Theory for Assessing Student Understanding - Robert Mislevy, ETS
10:00 - 10:30 Coffee
10:30 - 12:30 Uses and Misuses of IRT in Language Testing
   1. Exploring Inter-Rater Reliability with Rasch Techniques - Tim McNamara and Raymond Adams
   2. Let's Not be Too Rash About Rasch - Pardee Lowe, Danielle Janeczewski and Eduardo Cascallar
   3. Partial Credit Scoring of Cloze-Type Items - Ali Aghbar and Huixing Tang
      Discussant: John de Jong - Some Common Misconceptions about IRT: Possibilities and Impossibilities in Applications to Language Testing
12:30 - 1:30 Lunch
   CHAIR: Elana Shohamy
1:30 - 3:30 Testing Specialized Abilities
   1. Testing Translation Ability - Charles Stansfield, Mary Lee Scott and Dorry Kenyon
   2. Evaluation of the Defense Language Aptitude Battery - John Clark and Danush Hooshmand
   3. Measuring Growth in ESL Reading - Shelia Brutten, Kyle Perkins and John Upshur
      Discussant: Mari Wesche
3:30 - 4:00 Coffee
4:00 - 5:00 Criterion-Referenced Testing
   1. Item Discrimination Indices in Criterion-Referenced Language Testing - Thom Hudson
   2. A Criterion-Referenced Language Testing Project - Antony Kunnan
      Discussant: R. F. Boldt
5:00 - 5:30 Poster Session: Overview (5 minute overview)
5:30 - 6:30 Poster Session: Display
   1. The Carleton Academic English Language Assessment: Linking Testing and Learning - Janna Fox
   2. Breakdown and Repair - Donna Ilyin
   3. A Conversation Analytic Perspective on Interaction in the Language Interview - Ann Lazaraton
   5. Simulated Oral Proficiency Interviews for Less Commonly Taught Languages - Charles Stansfield and Dorry Kenyon
7:00 Dinner
1991 LANGUAGE TESTING RESEARCH COLLOQUIUM
FRIDAY, MARCH 22
CHAIR: Alan Davies

8:30 - 10:00 Identifying Constructs in Data from Language Tests
1. Trait Development in an Intensive ESL Course for Adolescents: Grammatical Knowledge and Communicative Ability - Caroïne Turner and John Upshur
2. Second Language Constructs as Indexed by ACTFL Ratings and TOEFL Scores - R. F. Boldt
3. Can We Test L2 Reading Comprehension without Reasoning? - Isabel Berman
   Discussant: Grant Henning

10:00 - 10:30 Coffee

10:30 - 12:30 Validating Oral Tests
1. Errors in Predicting Oral Comprehensibility from TOEFL and TWE Scores, in Relation to Observed Skill Levels - Gerald DeMauro
2. Qualitative Validation of Two Oral Tests - Elana Shohamy, Daphna Shmueli, Claire Gordon
   Discussant: Dan Douglas

12:30 - 1:30 Lunch

CHAIR: John Upshur

1:30 - 3:30 Comparing Tests and Items
1. Expert Estimates of Test Item Characteristics - Gary Buck
2. The Use of Test Method Characteristics in the Content Analysis and Design of EFL Proficiency Tests - Lyle Bachman, Fred Davidson and Michael Milanovic
3. Differential Item Functioning on Two Tests of EFL Proficiency - Katherine Ryan and Lyle Bachman
   Discussant: Alan Davies

3:30 - 4:00 Coffee

4:00 - 6:00 Business Meeting
1. Becoming an organization?
2. Adopting testing standards?
3. Next year's location/theme?

7:00 Banquet
1991 LANGUAGE TESTING RESEARCH COLLOQUIUM
SATURDAY, MARCH 23
CHAIR: Charles Stansfield

8:30 - 10:00 Testing ESP
1. SPEAK and CHEMSPEAK: Measuring the English Speaking Ability of International Teaching Assistant in Chemistry - Dan Douglas and Larry Sclinker
2. The Effect of Academic Discipline on Reading Test Performance - Caroline Clapham
Discussant: Charles Alderson

10:00 - 10:30 Coffee

10:30 - 12:30 Applications of Language Assessments
1. Norms Applicability of English Achievement Tests in K-12 ESL/BE - Fred Davidson
2. Can Test Bias Be Claimed When Language Proficiency is Inadequate - Alan Davies
3. The Role and Limitations of Self Assessment in Testing and Research - Doreen Ready-Morffitt
Discussant: Kathleen Bailey

12:30 - 1:30 Lunch
CHAIR: Jacqueline Ross

1:30 - 2:00 Poster Session: Overview (5 minutes each)

1. Constructed-Response Questions and Third World Teachers - David Carroll and Johanna Kowitz

2. Recent IRT Software Developments - John deJong

3. Assessing The Academic Achievement of Language Minority Students - Margo Gottlieb

4. The Validation of the UCLA English as a Second Language Placement Examination (ESLPE): A Study of Latent Factor Structures - Toru Kinoshita

5. Considerations in the Development and Trialling of CASE - Michael Milanovic, D. Foll and N. Saville

2:00 - 3:00 Poster Session: Display

3:00 - 3:45 Standards in British Testing: A Study of the Practice of British Examination Boards in EFL/ESL Testing - Charles Alderson and Gary Buck
Discussant: Liz Hamp-Lyons

TRAIN/BUS TO TESOL
Paper Abstracts Alphabetically Arranged
Aghbar, A.  
Indiana University of Pennsylvania  
Tang, H.  
University of Pittsburgh  

PARTIAL CREDIT SCORING OF CLOZE-TYPE ITEMS

In second language acquisition research it has been well accepted that acquisition of target language (TL) forms does not take place in an all-or-none fashion; rather, learners follow a certain systematic, developmental route in gradually approximating the TL forms. Yet popular modes of standardized testing in ESL, characterized by dichotomous scoring, have failed to measure this developmental process. The widely used cloze test, for example, generally characterizes responses as right or wrong and does not distinguish among "incorrect" responses which may in fact indicate varying levels of TL mastery. The purpose of this study is to: A) explore ways of scoring cloze-type responses on a polychotomous scale by assigning partial credits to "incorrect" responses, B) use the Rasch Item Response Theory (IRT) in its partial credit form to analyze and calibrate a set of vocabulary items, and C) see whether IRT can provide a more accurate assessment of language ability than the traditional methods.

At the piloting stage a 50-item cloze-type test of collocational vocabulary was administered to 100 ESL college students. From this, we will select a subset of 25-30 "best" items based on item difficulty and discrimination indices. We will then administer this subset to 300 ESL college students. An appropriate package for IRT analysis will be run on the results of the test to obtain ability estimates for each person and difficulty estimates for each step in each question as well as fit statistics and other relevant information. Further analyses will be done to examine the validity of our method. Implications for ESL teaching and testing will be considered.
Alderson, J. C. and Buck, G.
Lancaster University

STANDARDS IN BRITISH TESTING: A STUDY OF THE PRACTICE OF BRITISH EXAMINATION BOARDS IN EFL/ESL TESTING

It has often been asserted (e.g., Alderson, 1987, Bachman et al, 1989, Alderson, 1990) that the British tradition in language testing is substantially different from the North American tradition. Whereas in North America it is considered normal practice to examine the statistical and psychometric properties of tests, in Great Britain far more emphasis seems to be placed upon test content. However, it would appear that there has been no systematic study of this somewhat stereotypical view. The purpose of the present study is to begin an investigation of this issue by examining the current practices of British examination boards.

A questionnaire was constructed to establish: a) whether examination boards have a set of standards to which they adhere; b) what procedures they follow for estimating test reliability; and c) what procedures they follow to ensure test validity.

The questionnaire was sent to all examination boards listed in Carroll and West (1989), and also to the Schools Examination and Assessment Council. Where necessary, follow-up inquiries were made requesting clarification or expansion.

Results indicate that concern with the lack of statistical analysis of British language tests seems to be well founded. Although there were notable exceptions discussed in the paper, it is clear that the standards and procedures followed by most British examination boards fall far short of those recommended in basic introductory language testing or psychometric textbooks.

The paper discusses the implications of these findings for the professionalism of British language testing, and relates the outcomes to the incipient debate on the need to establish clear and appropriate professional standards for language test construction.
Bachman, L.
University of California, Los Angeles

Davidson, F.
University of Illinois

Milanovic, M.
University of Cambridge Local Examinations Syndicate

THE USE OF TEST METHOD CHARACTERISTICS IN THE CONTENT ANALYSIS
AND DESIGN OF EFL PROFICIENCY TESTS

The literature on language testing abounds with claims that content
considerations are essential to both the design and validity of language
tests. Such claims reflect common practice in the field where language tests
are typically developed and written on the basis of test content specifications,
and where descriptions of content are often cited in support of validity.
Furthermore, the measurement literature on validity is quite clear
that although content relevance is of limited use as the sole evidence for
validity, it does provide an important type of evidence in the investigation
of construct validity. All of this suggests that if we pay careful attention
to the specification of test content in the design and development of language
tests, we will improve the likelihood that the interpretations we make on the
basis of test scores will be valid.

This study extends research reported earlier on the use of content
analysis in the comparison of two different EFL proficiency test batteries, to
the design and development of a single EFL proficiency test battery. The
purposes of this study were: 1) to describe the content of multiple forms of
an EFL proficiency test in order to investigate their content comparability
and the relationships between test content and item statistics, and 2) to
provide feedback to the test developer for use in the revision of test content
specifications and in the ongoing analysis of test content to assure content
comparability across forms.

A rating instrument, based on Bachman's (1990) frameworks of
communicative language ability and test method facets, that was developed and
revised on the basis of earlier research, was used by five expert raters to
quantify their judgments about the characteristics of test items and passages
in multiple forms of the First Certificate in English (FCE), Papers 1 (Reading
Comprehension), 3 (Use of English), and 4 (Listening Comprehension). G-theory
was used to estimate the consistency of composite ratings which were used as
the basis for further analyses. Differences in composite ratings across forms
were used to examine their comparability, while multiple linear regression
analysis was used to investigate the relationship between test content and
item statistics. The implications of the results of this study for the use of
content analysis in the design and quality control of future FCE forms is
discussed. The results are also discussed in the context of past studies in
the field, with implications suggested for language test development practice
and future research.
CAN WE TEST L₂ READING COMPREHENSION WITHOUT TESTING REASONING?

The English Reading Comprehension (ERC) subtest of Israel's College Entrance Examination is currently constructed according to detailed specifications based on reading theory and linguistic features (e.g., extracting main ideas, recognizing cohesive devices) as well as the statistical parameters of the test. The correlation between the ERC and the Hebrew Verbal Reasoning (HVR) subtest has recently risen significantly enough to call into question their relative independence.

Despite tacit acceptance of Thorndike's 1917 statement that "reading is reasoning," several research studies are currently under way to investigate both the content and construct validity of the ERC and HVR subtests. In the first study, the ERC subtest given in October 1990 will be classified according to the above-mentioned specifications and then correlated with the HVR subtest. Factor and cluster analysis will be performed. An attempt will be made to classify ERC items in relation to their correlation with HVR. Given that hierarchical classification can be established, a new Table of Specifications may be constructed which places constraints on the use of items that correlate too highly with HVR.

In a second study, an experimental English subtest in which the majority of the items are linguistics-based, requiring little "reasoning," will be constructed, piloted, and correlated with the operational ERC and HVR scores for the same population.

The various studies to date, their statistical data, and their implications for future English Reading Comprehension test construction will be the focus of this paper.
SECOND LANGUAGE CONSTRUCTS AS INDEXED
BY ACTFL RATINGS AND TOEFL SCORES

Three numerical section scores result from the Test of English as a Foreign Language (TOEFL) program of ETS. The sections are Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary. Some educators suggest that the scores would be more meaningful if indexed by brief descriptions of the behavior levels implied. The descriptions would refer to behavior that occurs apart from the test-taking situation.

Our approach was to use the American Council of Teaching of Foreign Languages (ACTFL) descriptors. ACTFL has developed descriptors in paragraph form for the domains of interest, i.e., listening, writing, and reading. By indicating the appropriate ACTFL descriptor level, ESL instructors rated students' proficiency. The basic data of the study, then, are the students' TOEFL scores and a record of who rated each student and the ratings given.

Results presented will be from analyses as follows: First, we developed typical descriptive data. Second, we used several schemes to quantify the verbal ratings and compare the results. Third, we obtained several estimates of the reliabilities of the ratings. Fourth, we obtained evidence relating to the differential validity of measures from the three domains. Finally, the distributions of ratings at levels of the TOEFL section scores were developed. These latter distributions will be critical in defining the verbal interpretations of the TOEFL scores.
Brutlen, S. R.
Perkins, K.
Southern Illinois University
Upshur, J.
Concordia University

MEASURING GROWTH IN ESL READING

This paper reports on one part of a research program directed towards the description and measurement of growth in ESL reading comprehension. The aim of this study is to identify subskills assessed by reading comprehension test items and to assess the relation of these skills to reading proficiency in English as a second language. Subjects are university students taking courses in ESL; proficiency is defined operationally within the teaching program; test item data will come from an institutional administration of the TOEFL.

A panel of researchers will use various taxonomies of reading comprehension skills that reflect pedagogically significant aspects of reading in an examination of each TOEFL reading comprehension item. They will, independently, determine the reading skills that the items appear to assess. Items with agreed upon categorizations will be partitioned into subtests. Next the subtests will be submitted to latent trait analysis to identify misfitting subjects and items. These will be removed from the data matrix.

Subtest scores will be regressed on proficiency to estimate a measure of growth. A second, simple analysis will examine the correlations of subtest items and proficiency.

Results will be discussed in relation to the following questions:

1. To what extent are skills underlying reading test items identifiable?

2. To what extent are pedagogically important subskills represented in a standard reading test?

3. What subskills are most indicative of growth in reading?
Buck, G.
Lancaster University

EXPERT ESTIMATES OF TEST ITEM CHARACTERISTICS

The purpose of the present study is to investigate the extent to which "experts" can examine test items and predict their characteristics. The items used were from an ESL listening test, with 33 short-answer comprehension questions on a short narrative text. Twenty of the items were constructed to measure lower-level processing, defined as understanding clearly stated information; 13 were constructed to measure higher-level processing, defined as requiring inferences based on that information. There were 20 experts, all applied linguists with an active interest in second language testing or second language comprehension. The experts were asked to rate each item on two characteristics: (i) whether the item was testing lower- or higher-level processing, and (ii) the relative difficulty of the item. The tests were then administered to 254 students at a number of different colleges throughout Japan. Results showed strong agreement between experts with items written to test clearly stated information, but less agreement on items written to test inferences. In the case of item difficulty there was almost no agreement at all between experts; correlations between them were generally low, as were those with actual difficulty measures. Even when the 20 expert ratings were pooled, correlations with the actual difficulty measures were still low, suggesting that experts really do not know what makes test items difficult.
C. Clapham, Lancaster University

THE EFFECT OF ACADEMIC DISCIPLINE ON READING TEST PERFORMANCE

There have been several studies into whether background knowledge affects test scores in university language proficiency tests, but none have produced conclusive evidence either for or against the inclusion of ESP tests at this level.

At LTRC 1990, the author described a pilot study in which students took two reading comprehension tests, one within and one outside their academic field of study. Repeated measures analysis of variance was used to see whether students' scores were higher when they took tests within their own subject area. From this pilot study there was little evidence that they were, and indeed in some cases individual students did better at tests which were outside their academic discipline. This seemed to support the case for omitting the ESP element from university proficiency tests. However, these results have to be viewed with caution as the number of students in the study was small and no account was taken of the level of students' language proficiency.

A follow-up study on several hundred students has now been carried out. As in the previous study, students took two of the academic reading modules from the International English Language Testing System (IELTS) test. This time, however, in the ensuing examination of results, language proficiency, academic area, and background knowledge were all taken into account, and the data was analyzed using analysis of variance and covariance.

This paper describes the results of this study and discusses their implications for the construction of future proficiency tests.
EVALUATION OF THE DEFENSE LANGUAGE APTITUDE BATTERY (DLAB)

Defense Language Aptitude Battery (DLAB) is the major vehicle used by the Department of Defense to screen and identify students who have the potential to study a foreign language and succeed in meeting established proficiency standards as measured by the Defense Language Proficiency Tests (DLPTs) in a large number of languages. In addition to their use in selecting students for studying foreign languages in general, scores on this test are also used to assign selected students to different foreign language programs. This assignment is made through recommended minimum DLAB scores for each of the four categories of languages (languages taught at government agencies have been grouped into four categories based on their level of difficulty for native English-speaking learners of those languages) such that students with higher scores are assigned to more difficult languages.

DLAB consists of 126 multiple-choice items in four parts. The first part is designed to obtain biographical and attitudinal information. The second part measures the ability to recognize stress patterns in utterances made in an artificial language. The third part consists of four sections and measures the ability to learn and apply various grammatical rules to interpret short phrases and sentences in an artificial language. For example, in the first section of this part of the test, the examinee reads in the test booklet two rules that deal with nouns and adjectives. Then, in each item, the examinee reads a short phrase or sentence in English and hears on the tape four utterances, one of which is a correct translation of that phrase or sentence. The examinee's task is to identify the correct translation through applying the grammatical rule provided in the section. The fourth part deals with concept formation procedures and is designed to measure the examinee's ability to learn and apply conceptual features to interpret concepts appearing in new configurations in each item. It takes about one and a half hours to administer this test.

This paper reports the results of a study which seeks to evaluate DLAB to identify the strengths and shortcomings of the test in predicting language learning success across four languages (Russian, Arabic, Spanish, and German), each representing one of the four categories of languages discussed above. The major focus of the study involves an evaluation of various statistical characteristics of individual DLAB items.
NORMS APPLICABILITY OF ENGLISH ACHIEVEMENT TESTS IN K-12 ESL/BE

The Illinois State Board of Education (ISBE) Evaluation Section collects data on K-12 programs serving language minority students (LMSs) in Illinois. After summarizing findings for ISBE LMS data collection in recent years, this paper reports on some specific research related to the 1987-88 and 1988-89 end-of-year data collection. In those years, the ISBE also administered nationally standardized tests to a sample of K-12 ESL/Bilingual Education (BE) students in Illinois. The specific research reported here relates to norms applicability: the question of whether it is statistically appropriate to administer English language achievement tests normed on native English speakers to nonnative speaking students enrolled in ESL/BE programs. Comparability of variance, comparability of reliability, and comparability of subtest trait structure (via factor analysis) are discussed. Results indicate that, on the whole, these tests were statistically appropriate for these groups, even though the national norming did not include ESL/BE students. Implications of such a finding will be the focus of the audience discussion.
SOME COMMON MISCONCEPTIONS ABOUT IRT: POSSIBILITIES AND IMPOSSIBILITIES IN APPLICATIONS TO LANGUAGE TESTING

The history of Item Response Theory (IRT) can be traced back to early efforts of modeling nonphysical phenomena at the beginning of this century (Binet & Simon, 1916; Thurstone, 1925). It is only in the second half of the century, however, that mathematically sound models were developed, and only in the past decade has the application of these models been made feasible by the wider availability of computing power. The last decade will show a substantial reduction in costs of applications of IRT, which will effect proliferation at an unheard of scale.

This popularity of IRT will undoubtedly enhance the danger of misapplications of IRT and misinterpretations of its results in the same way as this has occurred with, for instance, Factor Analysis.

This contribution will discuss some theoretical notions involved in IRT and subsequently deal with erroneous applications of IRT based on misinterpretations of the underlying theory.

First it will be argued that many objections raised against IRT are based on disregarding that IRT is a theory. The Rasch model is based on a theory of what measurement is and the essential properties involved in any measurement procedure. Moving away from these a priori requirements into the realm of observable data, psychometric models have been proposed that lack these theoretical underpinnings.

In the second part, issues such as unidimensionality, scalability, sample-free item calibration, test equating, guessing, and item-discrimination will be dealt with both from a theoretical and a practical point of view. Each theoretical notion will be exemplified with practical examples based on real data taken from different studies in the field of language testing. Particular attention will be paid to the interpretation of analyses performed on actual test data, and the type of decisions that can be based on these interpretations.
Davies, A.
University of Melbourne

CAN TEST BIAS BE CLAIMED WHEN LANGUAGE PROFICIENCY IS INADEQUATE?

This paper describes an evaluation of an Australian statewide (N=60K+) primary school program of Basic Skills Tests of Numeracy for potential bias against children of non-English speaking background (NFSB) and raises the question of whether these tests are de facto language tests. The Basic Skills Tests (BST) of Numeracy are based on the New South Wales curriculum designed for all 7- and 10-year-old children, about 15% of whom are NESBs. This curriculum reflects current ideas of mathematics education, which emphasizes the centrality of language in the development of numeracy, an emphasis sometimes referred to by the slogan "No Naked Numbers!"

Although the NESB results on the 1989 Basic Skills Tests of Numeracy were consistently lower than were those of the English-speaking children, it is concluded that there is no evidence of systematically unfair bias against children of non-English speaking background (NESB). Bias is interpreted as being a deliberate misrepresentation of the skill in numeracy that NESB children are capable of achieving. Examples of three types of potential cultural bias discussed are: different conceptual structure, local knowledge, and unnecessary language difficulty, which it is concluded is the major problem.

Since there is no evidence that NESB children are intrinsically less able than English-speaking children in numeracy skills, it is therefore concluded that their BST performance is an example of underachievement caused by their overall inadequate English language proficiency.

Implications are drawn for practice in terms of test revision (for example, by reducing language difficulty in the test items) and fairer distribution of resources, and for theory in terms of the relation between language and culture in language test construction.
ERRORS IN PREDICTING ORAL COMPREHENSIBILITY FROM TOEFL AND TWE SCORES
IN RELATION TO OBSERVED SKILLS LEVELS

Overview

Analyses of the Test of Spoken English (TSE) show that Overall
Comprehensibility scores are well predicted (r>.9) by diagnostic ratings of
oral grammar, pronunciation, and fluency skills (DeMauro, 1988). However, at
higher skills levels, these diagnostic ratings under-predict the observed
comprehensibility rating.

Perhaps there are nonlinear components of Overall Comprehensibility in
relation to the other measured oral skills. For example, increased fluency
may not be uniformly related to increased Overall Comprehensibility throughout
the range of skills.

Research Question

This finding raises the question of what additional information is
provided by the Overall Comprehensibility rating that is not available from
measures of other oral and written English skills. The technical analog of
this issue is whether the Overall Comprehensibility rating has differential
discriminant validity throughout the range of oral skills.

Sample

In May 1990, 850 examinees took the TSE, the Test of Written English
(TWE) and the TOEFL. From their scores, prediction models will be developed
for Overall Comprehensibility using Listening Comprehension, Structure and
Written Expression, Vocabulary and Reading Comprehension, English essay-
writing skills, oral grammar, oral pronunciation, and oral fluency as
independent variables.

Analyses

Multivariate regression (general linear and step-wise) will predict
Overall Comprehensibility scores from TOEFL section scores and TWE scores.
Mean regression residuals will be plotted and analyzed by observed Overall
Comprehensibility score intervals.

Conclusion

Regression residuals will be related to the focus of the Overall
Comprehensibility rating scale to discern if certain types of component
skills are more prominent at different score levels, and what this means to
the discriminant and convergent properties of Overall Comprehensibility.
Douglas, D.  
Iowa State University  
Selinker, L.

SPEAK AND CHEMSPEAK: MEASURING THE ENGLISH-SPEAKING ABILITY OF INTERNATIONAL TEACHING ASSISTANTS IN CHEMISTRY

Research question: Is there a measurement advantage in using a fieldspecific test of English over a general test to evaluate the speaking skills of international teaching assistants?

Rationale: In a pilot study, Douglas and Selinker (1990) compared the performance of international teaching assistants in mathematics on the general SPEAK test and a field-specific version. They found statistically significant differences in grammar and fluency subscores, and a number of rhetorical differences in the responses. Other researchers (e.g., Clapham 1990, and Smith 1988) have found varying degrees of difference between general and field-specific measures. The present study refines the field-specific testing procedure and explores the relationship between speaking proficiency and field-specific teaching performance.

Method: Twenty-two prospective international teaching assistants in chemistry and biochemistry took the standard SPEAK test and a chemistry version back-to-back, in alternating order. The tests were scored by the regular, trained SPEAK raters, and the results compared with a criterion measure, a teaching performance test.

Analysis: Correlational procedures were used to investigate any advantage of either test in predicting performance on the criterion measure. In addition, transcripts of the subjects' responses on the two tests were analyzed as discourse to explore possible sources of variation indicated by the statistical analysis.

This paper will report on the results of the analyses, and begin to elaborate a theory of field-specific language testing.
Thom Hudson
University of Hawaii at Manoa

ITEM DISCRIMINATION INDICES IN CRITERION-REFERENCED LANGUAGE TESTING

Item analysis of criterion-referenced tests (CRTs) presents several practical problems. Item discrimination indices such as point biserial correlations are of limited informativeness if score distributions are narrow. When no prior defined mastery group is available, application of the CRT item difference index is not possible. Additionally, in settings which have relatively small numbers of examinees, item-response theory (IRT) methods will not yield stable estimates. Likewise, in many language programs either IRT computer programs are unavailable or the results of IRT analysis will be uninformative to those involved in test development. This study examines the relationship of three-item discrimination indices to IRT results in order to provide testers with information which will be useful in contexts in which no identifiable mastery group is available or when IRT analysis is inappropriate.

Three indices which indicate item discrimination at the cut score when no prior defined mastery group is available are compared to IRT results on data from two types of language tests: The indices are the Phi-Coefficient ($\phi$), the B-Index, and the Agreement Statistic. The two types of language test data are: (1) a multi-level general proficiency test, which is examined to determine the extent to which the indices reflect mastery of levels, and (2) achievement tests from the English Language Institute at the University of Hawaii, which are examined to determine the extent to which the indices reflect mastery of course goals. Implications for CRT development and analysis are presented.
Kunnan, A.
University of California, Los Angeles

A CRITERION-REFERENCED LANGUAGE TESTING PROJECT

The primary purpose of ESL placement examinations in North American universities has been to diagnose the level of English ability of nonnative English-speaking students, as a basis for making decisions about placement into specific ESL classes or exemption from such classes. With few exceptions (e.g., Brown, 1990), most ESL placement examinations are currently developed on the basis of norm-referenced (NR) measurement principles—either classical true-score theory or item-response theory. However, there is considerable research in the measurement literature, as well as at least one study in the language testing literature (Hudson & Lynch 1984), indicating that criterion-referenced (CR) measurement principles are more appropriate to tests that are based on specific content objectives and levels of achievement, and that CR approaches provide information that is relevant to placement decisions, information that cannot be obtained with NR approaches.

This study reports on the development and research of an ESL placement test using CR testing procedures at a public university in southern California. The items for the test were developed using the following procedures: definition of subskills, choice of appropriate item format, the piloting of test items, and initial NR measurement statistics. The CR dependability of the test scores across ability groups and of placement decisions across placement levels was then analyzed in the framework of G-theory, with a software program called GENOVA (Crick & Brennan 1983). Dependability indices indicated that the test is not equally dependable for all ability groups, while agreement indices indicated that the dependability of placement classification decisions differed considerably across the range of placement levels. Further research that will be done to investigate whether the test measures the same abilities at different levels includes exploratory and confirmatory factor analyses.

In closing, the study discusses the usefulness of a CR language testing project for placement officials and makes suggestions for standard setting, cutoff scores, and placement decisions.
LET'S NOT BE TOO RASH ABOUT RASCH

This presentation will address the theoretical and practical implications of Rasch analyses (Bigscale by Wright, Linacre, and Schultz 1990) for the scaling of two full-range ILR reading proficiency tests—the first in Dutch, the second in Norwegian. These tests cover the ILR reading proficiency scale, which ranges from 0 through 5 with pluses at all levels save 5, and are designed to assess general reading proficiency. The Dutch test contains 600 to 650 items, the Norwegian 114. Being criterion-referenced tests with multiple-decision points, ILR examinations of this type differ markedly from tests with a unimodal decision point common elsewhere.

The nature of other data (e.g., concurrent validity with other tests, with exemplary performances of known candidates in the classroom, and the nature of analyses besides those obtained by a Rasch approach) will be discussed to show the variety of factors which can affect scaling, in ways which are different from those suggested by Rasch analyses. Several aspects to be discussed include the importance of the size and distribution of the field-test population on Rasch statistics, the effect of including or removing persons and items from the analyses, and the nature and causes of misfitting items. Implications for the interpretation of the Rasch model under similar conditions, as well as the strengths and limitations of the Rasch analysis for the scaling of full-range reading proficiency tests will be examined.

Finally, a case will be made for a weighted versus an additive scaling of this type of general reading proficiency test.
McNamara, T.  
University of Melbourne

Adams, R.  
Australian Council for Educational Research

EXPLORING INTER-RATER RELIABILITY WITH RASCH TECHNIQUES

The increasing use of subjective assessments to measure performance on realistic written and spoken tasks in language tests has led to a corresponding need to establish the reliability and validity of such assessments. A difficulty arises in these situations because raters contribute an additional source of variation to the measurements. This variation can be considerable, and cannot be ignored. Current procedures for controlling this additional source of variation typically include the expensive multiple rating of scripts of tapes.

Recent developments in multi-faceted Rasch measurement (Linacre, 1989) provide improved mechanisms for investigating inter-rater consistency, rater “harshness” and the like. In this paper we describe the analysis of scripts from 100 candidates taking the International English Language Testing System (IELTS) test, a test of English for Academic Purposes, which were marked by two raters. The analysis illustrates how multi-faceted Rasch measurement can be used to examine inter-rater consistency, differences in rater harshness, differences in the manner in which raters use the available grades on the rating scale and the effect that between rater variation has on the measurement of individual candidates.

The use of the model to determine the amount of multiple marking required to achieve stable ratings of candidates is considered, illustrating how multi-faceted measurement techniques can be used in the development of cost-saving assessment practices without sacrificing reliability.
Doreen Ready-Morfitt  
University of Ottawa  
Second Language Institute  
600 King Edward Avenue  
Ottawa, Ontario  
KIN 6N5  

THE ROLE AND LIMITATIONS OF SELF-ASSESSMENT IN TESTING AND RESEARCH

The research is being conducted at a university second-language institute which annually tests for placement of approximately 500 students in the English Second Language (ESL) program, and 1,000 students in the French Second Language (FLS) program. Self-assessment has been used as the only means of placement since 1984, given the need for initial placement in courses by mail registration, and based on early positive results reported by LeBlanc and Painchaud (1984).

In 1987, the focus of placement changed with the introduction of a comprehension-based instruction program designed to bring beginning second-language learners to an intermediate level of proficiency. The program consists of four instructionally linked 3-credit courses of which only the two higher level courses are credit courses. Such a program depends on the homogeneity of classes at each level, so it is essential that students be properly placed.

In order to test the limits of self-assessment as a placement tool in this program, a standardized placement test was developed by the ESL department, based on comprehension tasks, content, and objectives from each of the four levels. The test was validated on the basis of each student's subsequent performance in the classroom.

After determining student placement using this instrument, a study was conducted comparing the results of the two placement methods. At the two lower levels, approximately 75 percent of the students were misplaced on the basis of placement by the self-assessment questionnaire. Results at higher levels were better, but in no case was the misplacement rate less than 40 percent. Since the self-assessment instrument clearly demonstrated a sufficient range of tasks to discriminate at lower levels, another approach was taken to try to increase the accuracy of initial placement through self-assessment. Mail-out instruction to students included a (radio) listening and cloze task in their second language and instructions as to how to interpret their performance. Students were directed to fill in the self-assessment questionnaire only after completing these tasks.

The results reported will be those relate to the success of initial placement of the September 1989 incoming students based on the changes to the self-assessment procedure. The paper will discuss the implications of these findings in the context of the role of self-assessment in testing and research.
Ryan, K. E.
University of Illinois

Bachman, L.
University of California, Los Angeles

DIFFERENTIAL ITEM FUNCTIONING ON TWO TESTS OF EFL PROFICIENCY

Detecting items which do not function the same for specific subgroups (differential item functioning [DIF]) has been the focus of numerous investigations on tests of educational achievement (Williams 1985; Linn, Levine, Hastings & Wardrop 1981, McPeek & Wild 1986, Ryan 1990, Zwick & Ericakan 1989). These studies have primarily examined item performance differences between, respectively, males and females, blacks and whites, and Hispanic and whites. While the origins of detecting DIF in educational measurement are based on bias issues, these investigations have also suggested that these differences may not be "bias." Rather instructional and curricular differences between these groups may impact item performance (Linn & Harnisch 1984). In the L2 testing context, L1 background as well as curricular and instructional issues are of interest. Although the effects of differences in L1 and background knowledge on L2 test scores are well documented in the language testing literature (e.g., Swinton & Powers 1980, Alderman & Holland 1981, Oltman, Stricker & Barrows 1988, Alderson & Urquhart 1985, Hals 1988), the extent to which these differences lead to DIF has not been widely investigated.

This study examines the extent to which items function differentially for test takers of equal ability from different L1 and curricular backgrounds on the Test of English as a Foreign Language (TOEFL) and the First Certificate of English (FCE). Item responses of approximately 1,600 subjects in eight different countries that were collected as part of the Cambridge-TOEFL Comparability Study will be analyzed using the Mantel-Haenszel (MH) procedure and other indices proposed by Holland (1985) to detect DIF on the TOEFL and FCE multiple-choice items. Analyses will be conducted based on gender, Asian and non-Asian background, and individual country pair-by-pair comparisons where sample size is adequate. Items that are identified as differentially functioning will be content-analyzed for any patterns.
Shohamy, E., Shmueli, D., and Gordon, C.
Tel Aviv University

CONTENT AND CONCURRENT VALIDATION OF A
DIRECT VS. SEMI DIRECT TESTS OF ORAL PROFICIENCY

Recently, the Center of Applied Linguistic (CAL) has developed a series of semi-direct oral tests for the less commonly taught languages (Stansfield and Kenyon, 1988). These tests are to be used in situations where trained oral interview testers are not available. The tests are termed "semi-direct" since they are structured, yet contain communicative features which elicit a range of oral interactions in discourse. A semi-direct test consists of tasks such as giving directions, descriptions, topical discourse and simulations. The test taker uses the language by employing visual as well as aural stimuli and the oral language is then rated with the ACTFL guidelines.

In the process of development of a semi-direct test the concurrent validity of the test is examined against the scores which the same test takers obtain in a face to face Oral Proficiency Interview (OPI). The average concurrent validity which was obtained in the different languages for which the test was developed was $r = .93$, which lead the developers to conclude that the semi-direct tests are valid against the OPI criterion.

While correlations obtained from concurrent validations studies provide important information in test development, it has been claimed that such information is not sufficient as there is a need to examine qualitatively whether two tests measure the same things in regard to the cognitive processes involved in processing the language on each of the tests and/or the type of language which is elicited in response to the different testing tasks. (Grotjahn, 1986).

This paper will report on a qualitative validation study which investigated the concurrent validity of a semi-direct test in Hebrew (Shohamy, Gordon, Kenyon and Stansfield, 1989) in relation to an OPI Hebrew test taken by the same test takers. It will show that in spite of the high correlations obtained between the two types of tests, they differed in terms of the type of language samples which were elicited.

The data analysis was performed in two phases. In the first phase 6 oral language tests (3 of each oral test type) were analyzed and categories were identified. The categories included grammatical structures, morphemes, syntactic structures, pronunciation, cohesive elements, cohesion markers, speech functions, discourse markers, register shifts, pauses, topic familiarity, genre, vocabulary, error types and a number of para-linguistic features. In the second phase the frequency and variation of these categories in 30 oral tests (15 of each type) were compared.

Results of the comparisons, along with further analysis by proficiency levels, will be reported pointing to the specific categories in which the two tests differed. The relationship between the language samples obtained and the methods of elicitation will also be reported and discussed within the context of language variation theory.
AN EXAMINATION OF TRANSLATION ABILITY

The Center for Applied Linguistics has developed the Spanish into English Verbatim Translation Exam (SEVTE) for use by the Federal Bureau of Investigation for selecting language specialists who will serve as full-time translators and for determining the translation ability of special agents and support personnel whose regular duties include other functions. The exam consists of multiple choice and production items involving translation of phrases in sentences, whole sentences, and paragraphs. Thus, the SEVTE includes both indirect and direct measures of translation ability.

Translation skill level descriptions, modeled after Interagency Language Roundtable (ILR) level descriptions in speaking, writing, etc., were also developed in connection with the project. The translation level descriptions formed the basis for the scoring guides that were used to rate the production items.

Following a brief presentation of the exam format and scoring procedures, this paper will discuss the validation of the SEVTE. Scores on the test are analyzed along with scores on other measures. These include ILR Spanish and English ratings for listening, speaking, and reading; scores on other secure tests of translation ability developed by the federal government; and a holistic assessment of overall translation ability expressed on the ILR-like skill level descriptions. The results provide evidence that translation ability consists of two quite distinct traits: accuracy and expression. Accuracy represents the ability to render all of the propositional content of the source document into the target language document. Expression represents the ability to express oneself correctly in the target language when writing a translation. These results suggest that a single rating of translation ability, such as an ILR-like skill level description, is inadequate to explain the variance in translation ability. A suggestion is made as to how to revise the translation skill level descriptions to accommodate this concern. The results of the study provide a psychometric basis for understanding the construct of translation ability.
TRAIT DEVELOPMENT IN AN INTENSIVE ESL COURSE FOR ADOLESCENTS: GRAMMATICAL KNOWLEDGE AND COMMUNICATIVE ABILITY

In this paper we describe research to investigate the temporal order of attaining grammatical knowledge and of developing communicative ability in a communication-oriented L2 classroom. The project includes: i) development and validation of test batteries for communicative ability and grammatical knowledge appropriate to fifth grade students in a beginners' intensive ESL course, and b) development of multiple forms of these two batteries for a cross-legged time series investigation of language development.

Predominant conceptualizations of communicative competence (CC) today are componential. Second language teaching practices offer contrasting views of the process of language development. In one view, an increase in grammatical competency (GC) results in augmentation of CC. This view is implicit in courses that emphasize the teaching of language usage. In the contrasting view, an increase in CC through the effects of communicative strategies provides the requisite precondition for development of GC. This view is implicit in courses that emphasize language use. The relative adequacy of these two accounts of second language development may be evaluated using a single set of panel data.

Because instruments and specifications for instruments for the traits are lacking, the project requires development of defining measures. Tryout tests of both traits employing eight methods have been constructed. Following a feasibility study, test pairs will be investigated for convergent and divergent validity. If this investigation yields batteries of measures with sufficient reliability and validity, multiple forms will be constructed to provide the measures for the time series study which we have described.

In the paper we will elaborate upon the theoretical and methodological background of the project and will describe developments and findings to date.