A New Decade in Language Testing: COLLABORATION AND COOPERATION

TWELFTH ANNUAL LANGUAGE TESTING RESEARCH COLLOQUIUM

Metropolitan Club, San Francisco
March 3-5, 1990
MEMORIAL SERVICE
Michael Canale (1949-1989)

Monday, March 5, 1990, 3:45 p.m.
XII Language Testing Research Colloquium
San Francisco, California

Opening
Charles W. Stansfield
Center for Applied Linguistics

Contributions to an Understanding of Communicative Competence
Helmut Volmer
University of Osnabruck

General Contributions to Language Testing
Lyle Bachman, University of Language California at Los Angeles

Contributions to Language Testing in Canada
Margaret DesBrasay
University of Ottawa

Personal Tributes
Student
Janine Deen
Tilburg, Holland

Friend
Adrian Palmer
University of Utah

Colleague
Maurine McInerney
York University

Closing

Guest of Honor
Claire Trepanier
York University

Memorabilia will be on display following the service.
TWELFTH ANNUAL LANGUAGE TESTING RESEARCH COLLOQUIUM

SPONSORED BY
The English Language Institute Testing and Certification Division,
University of Michigan
and
The Testing Division,
Defense Language Institute Foreign Language Center
Presidio of Monterey

Conference Organizers
Liz Hamp-Lyons, The University of Michigan
Donna Ilyin, University of California - Berkeley
Dariush Hooshmand, Defense Language Institute

Proposal Reviewers:
J. Charles Alderson
Lyle Bachman
Andrew Cohen
Grant Henning
Dariush Hooshmand
Charles Stansfield

Tribute to Michael Canale organized by Charles Stansfield

Program
Registration; Hospitality
Local Arrangements
Printing; Local Co-host
A New Decade in Language Testing: Collaboration and Cooperation

CONTENTS

Program Introduction ........................................................................................................... 5
Daily Schedule ..................................................................................................................... 6-8
Abstracts (in chronological order of sessions) ............................................................... 9-37
Introduction

The Language Testing Research Colloquium is as active as ever as it holds its twelfth annual meeting. The large number of paper proposals testifies to the Colloquium's important role in the field of language testing research and in the calendar of language testers: despite determined efforts to squeeze a quart into a pint pot some excellent proposals had to be rejected. We hope you will agree that the quality and variety of presentations is such that none of those we offer you could have been omitted. Although we offered several innovations in presentation-types, these were not especially popular, and we shall be asking you yet again how you would like to see the Colloquium develop and change.

The theme of this year's Colloquium is 'Collaboration and Cooperation,' and the program clearly reflects the exciting trend in language testing research for groups to come together for common purposes, across countries, institutions and fields of interest. But even while language testers find ways to work together across large distances, the Colloquium still offers the opportunity to 'bash heads once a year, and have fun together' as Lyle and Buzz put it in their Introduction to the Tenth Annual LTRC. We hope this Colloquium will be a great opportunity to do both those things.

As we have worked toward this twelfth Colloquium we have never forgotten that this year language testers lost a wonderful colleague, friend and inspiration, when Michael Canale was killed in a climbing accident. The closing session of the Colloquium, planned by Charlie Stansfield, will be a memorial tribute to Michael. We think that Michael would be pleased if he could see from this program some of the directions language testing research is taking.

Liz Hamp-Lyons
Donna Flynn
Dariush Hooshmand
DAILY SCHEDULE

SATURDAY MARCH 3

8.30 - 9.30                      Registration & coffee
9.30 - 10.00                     Opening remarks
10.00 - 10.45 Gordon Hale & Rosalea Courtney
10.45 - 11.3 Grant Henning, Michael Anbar, Carl Helm & Sean D'Arcy
11.30 - 12.15 John H.A.L. De Jong & Grant Henning
12.15 - 1.15 Lunch
1.15 - 2.00 James Dean Brown
2.00 - 2.30 Elana Shohamy
2.30 - 2.45 Break
2.45 - 4.45 J. Charles Alderson, (Chair) John Foulkes, David Ingram & Caroline Clapham
7.00 Buffet Dinner FUN by RED RODGERS
8.30-9.30  NETWORK: Eduardo Cascallar, James Child, John Clark, Pardee Lowe, Thomas Parry & Charles Stansfield

9.30-10.00  Mary C. Spaan

10.00-10.30  Dan Douglas & Larry Selinker

10.30-11.00  Break

11.00-11.45  Tim McNamara

11.45-12.30  Thom Hudson

12.30-1.30  Lunch

1.30-2.15  Charles Stansfield, John Karl, Dorry Mann Kenyon & Dan O. Robertson

2.15-3.00  Eva Baker, Jean Turner & Frances Butler

3.00-3.15  Break

3.15-4.30  Lyle Bachman, (Chair) Fred Davidson, John Foulkes, J. Charles Alderson, John L. D. Clark, Bernard Spolsky & Charles Stansfield

4.30-5.30  Business Meeting (important business: please be there!)
MONDAY MARCH 5

8.30-9.15 Neil Anderson, Kyle Perkins, Andrew Cohen & Lyle Bachman
Construct Validation of a Reading Comprehension Test: Combining Sources of Data
Prompt Difficulty, Task Type and Performance Level in ESL Direct Writing Assessment

9.15-10.00 Sheila Prochnow

10.00-10.15 Break

10.15-11.00 Andrew Cohen
The Role of Instructions in Testing Summarizing Ability

11.00-11.45 J. Charles Alderson
Judgments in Language Testing

11.45-12.15 Doreen Ready
The Role and Limitations of Self-Assessment in Testing and Research

12.15-1.30 Lunch

1.30-2.00 Kyle Perkins & Sheila Brutton
A Comparative Analysis of Misfitting Items

2.00-2.30 Bernard Spolsky
TOEFL-The pre-history

2.30-3.00 Tibor Von Elek, Mats Oscarson & Fred Davidson
Implementation of National Composition EFL Tests in Sweden

3.00-3.30 Bronwyn Norton Peirce & Kathleen Troy
What the Autonomous Language Learner Can Teach Us About Assessment

3.30-4.45 Michael Canale:
A Tribute from LTRC
Note Taking and TOEFL Listening Comprehension

Abstract

The TOEFL Listening Comprehension section is designed to assess students' ability to comprehend spoken English as it usually occurs in an academic setting. A portion of the Listening Comprehension section consists of mini-talks, or short lecture segments of about 250 words each, with accompanying questions. Although note taking is not permitted, some people have suggested that allowing note taking in the mini-talk subsection would more closely simulate a typical classroom lecture situation, and thus increase the face validity of the test.

This study examines the effects of note taking on performance in TOEFL minitalks, in an experimental situation involving 560 students. Students are allowed to take notes for one set of mini-talks and not allowed to take notes for another set (with the order of note-taking and non-note-taking conditions counterbalanced across groups of students). The study inquires whether note taking affects mean performance, and also whether it affects test reliability, and relative standing of students. Also, using a questionnaire, the study ascertains students' views about being able to take notes in this situation and about their typical classroom note-taking habits, and it seeks to determine whether the students' questionnaire responses relate to the effects of note taking observed here.

Effects of two different types of note-taking situation are examined, one in which note taking is simply permitted, and one in which the students are urged to take notes (with the two conditions assigned to two different groups of students).

Performance data and questionnaire data are currently being analyzed, and results will be available well before the time of the Colloquium.
Grant Henning  
Educational Testing Service  

Michael Anbar  
The State University of New York at Buffalo  

Carl E. Helm  
R.W. Johnson Medical School  

Sean D'Arcy  
The State University of New York at Buffalo

A COMPARISON OF MULTIPLE-CHOICE AND OPEN-ENDED COMPUTERIZED  
ASSESSMENT OF ESL READING COMPREHENSION

The comparative functioning of multiple-choice and open-ended  
test items has long been a topic of interest to the testing community  
(Breland and Gaynor, 1979; Bridgeman, 1989; Ward, 1982). The proposed  
presentation will report results of a comparison of multiple-choice  
and open-ended test item performance in the assessment of ESL reading  
comprehension.

A pilot study was conducted with 44 university ESL students at  
the State University of New York at Buffalo. The subsequent main  
phase of the project involved the participation of a different sample  
of approximately 67 students at the same institution. Students  
responded to a set of 40 multiple-choice items drawn from a disclosed  
TOEFL reading component. An additional 40 open-ended questions  
designed to request the same information were also administered to the  
students within the controlled testing time. In addition, since the  
test was administered by computer, using the Computer Assisted  
Socratic Instruction Program (CASIP) authoring system designed by  
Anbar (1986), it was also possible to record response time for each  
student with every item.

Results will include mention of the comparative mean difficulties  
and discriminations of these multiple-choice and open-ended items.  
Furthermore, internal consistency reliabilities will be presented for  
the tests comprised of multiple-choice versus open-ended items. And  
results for a variety of scoring methods of open-ended items will be  
compared, including methods involving degree-of-correctness scoring  
and incorporating response time as an ability indicator. Preliminary  
analyses of pilot data suggest that adding degree-of-correctness  
information in the scoring of open-ended items may enhance reliability  
beyond levels commonly observed with the same number of multiple-  
choice type items.
TEST DIMENSIONALITY IN RELATION TO STUDENT PROFICIENCY

Dimensionality in foreign language tests is an important issue both from the linguistic and psychometric points of view. From the linguistic point of view, evaluation of the dimensionality of foreign language tests at different levels of proficiency may provide evidence to support models of foreign language acquisition. From the psychometric point of view unidimensionality allows for the interpretation of total test scores as sufficient statistics for student ability, whereas multidimensionality requires more complicated score reporting procedures (e.g. profiles) in order to reflect individual differences accurately.

The dimensionality of foreign language proficiency has also been an object of debate among scholars proposing the hypothesis of unifactorial general proficiency (e.g. Oller and Hinoftis, 1980) and those scholars proposing more complex hierarchical models (e.g. Vollmer, 1985). Studies using factorial designs have suggested different numbers of factors and also different sources for multifactorial solutions. For example, Upshur and Homburg (1983) suggest that dimensionality in language tests may be related to ability level, whereas Swinton and Powers (1980) have suggested a language background relation.

Previous studies of TOEFL dimensionality have reported differing numbers of underlying factors or dimensions depending on the methodology employed in analysis (Boldt, 1988; Hale, Rock and Jirele, 1989; Oltman, Stricker and Barrows, 1988; Swinton and Powers, 1980). Oltman, Stricker and Barrows (1988), using multidimensional scaling analysis, reported results suggesting that dimensionality may actually vary depending on the region of the difficulty-ability continuum under investigation.

The present study of language test dimensionality uses a random sample of 5,000 students from a population of TOEFL examinees which has been divided in three sub samples at three different levels of proficiency (low, mid, and high total score ranges). The study seeks to test the hypothesis that analysis of the same test response data with low-proficient examinees will indicate a greater number of dimensions underlying test performance than will analyses conducted with high-proficient examinees.

Several methods of analysis will be employed, including Rasch model fit analysis, latent structure analysis, and variations of the Bejar method. Explanations will be sought for all outcomes. The study will also constitute an example of international collaboration in language testing research.
James Dean Brown
University of Hawaii at Manoa

A COMPREHENSIVE CRITERION-REFERENCED LANGUAGE TESTING PROJECT

The English Language Institute (ELI) at the University of Hawaii at Manoa regularly offers seven courses in academic listening, reading, and writing. Over the last four years, the curriculum for each course has been extensively revised including thorough needs analysis, development of objectives, criterion-referenced tests, and materials, as well as improvements in teaching practices and regularly conducted formative evaluation procedures. This paper reports on one facet of the new curriculum: the criterion-referenced test development processes and results.

Each of the seven ELI courses has two forms of a criterion-referenced test written expressly to measure the objectives of that course. The two forms are regularly administered in a counterbalanced design such that half of the students take form A in the second week of classes while the other half take form B. At the end of the course, all students take the opposite form. Thus this testing project is relatively large in scale including 14 different tests administered before and after instruction for about 500 student enrollments per year.

While the objectives and resulting tests differ fairly widely in organization and form across the seven course, the processes involved in putting the tests in place were quite similar. Thus the initial item development, piloting and revision processes are described in general terms. Much more detail is provided about the results of the administrations of these CRTs during 1989. Descriptive and item statistics are presented for masters and non-masters (including difference indices) on each test. Dependability estimates [agreement, kappa, phi and phi(lambda)] are given, and evidence for the content and construct validity of the tests is also provided.

The discussion centers on the problems encountered in developing such a comprehensive testing program, then turns to the benefits which CRTs can provide for overall curriculum development.
Elana Shohamy
Tel Aviv University

A DIAGNOSTIC/FEEDBACK MODEL FOR ASSESSING
ACHIEVEMENTS AND PROFICIENCY IN FOREIGN LANGUAGES

The purpose of the paper is to describe a collaborative assessment project which is taking place between a University team in Israel and major schools teaching Hebrew in the USA and Canada. The main objective of the assessment is to provide schools with DIAGNOSTIC FEEDBACK about their students' level of achievements and proficiency at the end of six years of learning Hebrew, so as to facilitate improvement of teaching and learning in the schools.

The paper will describe the implementation and validation of an assessment model which is a ten-month process, that includes the following phases: Each school defines its objectives and contents for teaching Hebrew via a Table of Specifications, as a result of a workshop that the University team conducts with the principals and teachers of each school. According to each school's table, the University team tailor-makes language tests for each of the schools, in all the four language skills. The tests tap three main areas: a) school-based achievement - based directly on the objectives and content specified by the school; b) proficiency-based achievement - based on similar goals but on different material; and c) a core component which is based on 'agreed upon' language knowledge and is used to compare similar types of schools. The tests (and questionnaires regarding attitude, background, etc.) are administered at the schools and are then analyzed, diagnostically, according to a variety of sub-domains, in each of the skills, and in relation to achievement, proficiency and core knowledge. A report which includes all the above information is written by the assessment team and is delivered to each school in a lengthy meeting. The school, after discussing the results with the staff, students, parents, etc. write the last part of the report which addresses specific strategies they are planning to take based on the results. An instructional team then works with the school on strategies and methods for improvement accordingly.
Panel on the IELTS

Panel leader: J Charles Alderson, University of Lancaster
Participants: John Foulkes, UCLES,
David Ingram, Brisbane CAE, Australia
Caroline Clapham, UCLES and University of Lancaster

The British Council, the University of Cambridge Local Examinations Syndicate and the Australian International Development Programme have been collaborating since 1987 in the development of a new test of English for academic and general educational training purposes: the IELTS test. This test is a revision of the former ELTS test and the Revision Project has involved extensive international collaboration in test design, trialling, and analysis. Aspects of the Revision Project have been the subject of papers at the previous ILTRCs. Members of the project from the UK and Australia present papers on a number of aspects of the test battery which are thought to be either innovative or of interest to current concerns in language testing.
An example of international collaboration

This paper outlines the nature and development of the recently released ELTS test. This test is the result of a joint British-Australian project (with some Canadian input) to develop a new test for use with overseas students going to study in English-speaking countries. The test is a result of a three year project to revise the British Council and University of Cambridge Local Examinations Syndicate ELTS test. It measures candidates' general language ability and the skills needed for effective study or training. It includes modular tests in different subject areas, and test writing and speaking abilities directly. Test scores are reported according to bands of ability associated with performance descriptors.

This paper describes the final stages in the progress of the Revision project, paying particular attention to the preparation of the construction of assessment and training procedures for tests of writing and speaking. It concludes by discussing some of the issues which have faced the Revision Project, the decisions and the reasons for these, and highlights those aspects of the testing of English for Academic Purposes which need further research.

J Charles Alderson, University of Lancaster
John Foulkes, University of Cambridge Local Examinations Syndicate
David Ingram, Mount Gravatt College, Brisbane
The Role Of Grammar in EAP Testing

J Charles Alderson
University of Lancaster

Since the mid 1970s, there has been a growth of interest in the testing of English for Specific Purposes. In the UK, this has generally followed the model proposed by John Munby in 1978, and has involved making analyses of the candidates' "target situations": the domain of language use for which they require a language competence, and on which it is supposed that a test of such competence should be based. Specifications for such tests have typically included lists of enabling skills, text types and so on, and have tended to have a skill focus. One example of such a test is the ELTS, devised by the British Council and the University of Cambridge Local Examinations Syndicate. However, in the world of language teaching, ESP and, more recently, the "communicative approaches" have been followed by a concern that "grammar" should not be forgotten, since it is to be held at the heart of any language learner's competence. Interest has grown in devising a means of teaching grammar "communicatively", as is testified by an increasing number of textbooks and handbooks for teachers in the world of testing, also, interest has grown in testing such "communicative" grammar. The ELTS test has recently been revised, and one of the aims of the Revision Project was to examine the possibility of including a measure of "grammar" in the new battery. It was believed by consultant testing researchers that it was appropriate to attempt to test grammar separately from other skills. This paper reports the results of that attempt, and explains the reasons why it was decided not to include a test of grammar in the final battery.
Attempts at reliable innovation in listening

John Foulkes
University of Cambridge Local Examinations Syndicate

The IELTS test battery includes a test of listening which attempts to innovate in item types and, to some extent, in test content. Previous attempts by the University of Cambridge Local Examinations Syndicate to produce a variety of innovative listening tests in the First Certificate Exam, Paper Four, resulted in tests with less than optimal reliability. However, in the case of the IELTS test, not only were the item types innovative in the context of the ELTS tests (previously entirely multiple-choice) but the tests achieved very acceptable levels of reliability. In addition, the test format proved to be replicable, in that parallel forms of the test achieved high correlations with the original trial test. This paper will report on the nature of the test's design in terms of content and method, it will present the results of the various trials of the original test and its parallel forms, and it will speculate on what made it possible to produce an innovative test with high reliability.
Is ESP testing justified?

Caroline Clapham
University of Lancaster

The use of ESP testing is felt to have led to a significant advance in proficiency testing in Britain, but it brings problems in its wake. In the original version of the English Language Testing Service (ELTS) test for overseas students, for example, where there were subject modules in six different academic areas, many candidates had difficulty selecting appropriate modules. Even when they made a suitable choice, that choice was often far removed from their own discipline, and led to student frustration. Since doubling or quadrupling the number of modules is impractical, it has been argued by some language testers that ELTS should be turned into a general EAP test with no specific subject modules.

For the last three years there has been an international project to revise the ELTS test (now renamed IELTS). During this, the number of specific subject modules has been reduced to three, but the project is investigating the extent to which separate modules are justified for different groups of candidates. In an attempt to gauge the extent of advantage or disadvantage to students, candidates have taken two subject modules, one supposedly appropriate for their study intentions and one inappropriate. In addition performance on different modules has been compared with performance on a conflated module which is intended to produce only minimal disadvantage for any student.

The data has been analysed using item response theory, analysis of variance and different approaches to factor analysis. This paper reports on the results of this study and discusses its implications for proficiency testing.

ABSTRACT

After a long period of relative neglect, language aptitude assessment has once again become one of the areas of interest for some of those involved in the field of language testing. Some recent examples are two initiatives: (a) the Interagency Language Roundtable, promoting the development of collaborative efforts for research projects in the area, as well as co-sponsoring a national symposium on language aptitude assessment, which took place last September; (b) the Defense Language Institute currently working on a computer-delivered test for the assessment of the probability of success in language learning. At the same time, interest on this topic is the focal point of a proposed research plan at ETS, in collaboration with various government agencies, to develop a new approach to the problem of language aptitude assessment. It is expected that this approach will be complementary to the work currently under way at DLI, and will further our understanding of theoretical and practical issues from a componential perspective.

The various participants will share the work currently under way at their respective agencies. The representative from DLI will explain the state of the project at that institution, while other members of ILR and CAL will present their views on the problem of language aptitude assessment as well as the needs and possible applications at each agency. The framework for a new broad research plan on this topic will be presented by the ETS coordinator, and a detailed explanation of the theoretical and methodological underpinnings will be discussed. All components will be presented and input will be welcomed from those attending the network session.

To address the problem of language aptitude assessment current work includes the consideration of elements such as: cognitive style, motivational factors, prior language learning experience, etc. It is suggested that the inclusion of other cognitive processing factors, general cognitive and language specific learning strategies, as well as affective components, will improve the effectiveness of the assessment and its predictive value. Following Sternberg’s (1985) definition of a componential analysis methodology
for studying certain cognitive skills, procedures that draw upon both psychometric and cognitive methods will be used in testing the separate components. The underlying theory is evaluated in terms of five criteria: completeness, specificity, generality, parsimony and plausibility. The particular tasks which are suggested for inclusion in the analysis, satisfy four criteria, originally proposed by Sternberg and Tulving (1977) in a different context: quantifiability, reliability, construct validity, and empirical validity. In addition, tasks are decomposed into subtasks, in terms of the participation of a subset of information-processing components, which is a useful step in the analysis of the cognitive abilities in question. The rationale and method of task decomposition will be fully explained. It is expected, that with the full development of an instrument of this kind, a testing program could be established where the probability of achieving certain levels of language proficiency for the language learners in a given program could be established. Furthermore, in its full implementation, it is expected this cluster of task measures will allow the estimation of aptitude for different skills: listening, speaking, reading, writing. Another goal is to achieve equivalent predictive validity for various specific languages or language groups.
Mary C. Spaan  
The University of Michigan  

The Effect of the Prompt in Essay Examinations

In large-scale ESL writing assessment, only one sample is collected from each examinee, and usually only one prompt is assigned. It is crucial that the writing sample be representative of the writer's ability and that it will predict his performance on future writing tasks in academic or business settings. Yet research has shown that different types of writing tasks make different cognitive demands on writers, and may elicit different types of responses, which may not be assigned equivalent scores.

In this study, 88 students sitting for an ESL proficiency examination each wrote two essays. Two types of prompts, which made different cognitive demands, and required use of two different rhetorical modes, were used. Ninety percent of the students performed the same on the two essays.

The prompts will be analyzed in terms of cognitive demand, purpose, role, audience, content, and rhetorical specification (Vahapassi 1983, Purves et al. 1984). Next, the essays of the "inconsistent" writers will be matched with those of "consistent" writers at their overall proficiency level, and analyzed. Discourse level analyses will include measures of organization, cohesion, and framing (Connor 1988, Toulin 1958, Witte 1983). Sentence level features, such as total words, T-units, clause length, type/token ration, errors, will also be counted. Performance of "consistent" and "inconsistent" writers will be compared within and across proficiency levels on these variables using descriptive and correlational measures. If possible, the relative contribution of each of these variables to the overall holistic score will be assessed.

Such detailed analysis of response to different prompt types may prove helpful to developers of large scale writing tests, not only in designing appropriate and equivalent prompts, but also in reviewing their rubrics for scores on their writing scales.
PERFORMANCE ON GENERAL VERSUS FIELD-SPECIFIC TESTS OF SPEAKING PROFICIENCY

Users of the TSE/SPEAK have frequently suggested that the test is too general to be used as a valid measure of an ITA's ability to use English for teaching in fields such as mathematics, chemistry or physics. The test takers, and their supervisors, complain that if they were given a test of their abilities to talk about their major fields, they would do much better.

One study (Smith, 1988) was conducted to explore differences in performance on SPEAK and field-specific versions, and very little difference was found. However, it has been suggested (Douglas, 1989) that "contextualization cues" (Gumperz, 1978) present in the two testing situations may not have been sufficiently different to promote differential "domain engagement" (Douglas and Selinker, 1985), and consequently, differential interlanguage performance. The present project has been designed to explore that suggestion.

In our study, the SPEAK is given in the normal context of "evaluation of English speaking proficiency for prospective ITAs," while a field-specific version is being given in the context of "orientation for new ITAs in mathematics." The mathematics session will be conducted by a mathematics professor, and the stated purpose will be to give the new ITAs some experience at talking about mathematics in classrooms and to discover any problems they may have with teaching mathematics. All of the test instructions, as well as the items, have been revised in relation to this purpose. The test protocols will be scored by the usual trained, non-mathematics, raters. This procedure is being repeated in a chemistry domain.

The proposed paper will focus on a statistical and rhetorical comparison of performance on the SPEAK with that on the mathematics version. A validation study, involving evaluations from student in classes taught by the subject ITAs, will be suggested as a further step in the investigation.
T.F. McNamara
Department of Linguistics and Language Studies
University of Melbourne

ABSTRACT

ITEM RESPONSE THEORY AND THE VALIDATION OF AN ESP TEST
FOR HEALTH PROFESSIONALS

The issue of the empirical validation of constructs used in communicative language tests has been repeatedly raised in recent discussions. Skehan (1988) discusses the difficulty of finding appropriate settings (in terms of operationalization of constructs and subject size) in which to carry out the necessary research. Such a setting is provided by the introduction of a new ESP test for health professionals developed recently on behalf of the Australian Government.

The test includes subtests of speaking, writing, reading and listening, with provision for profession-specific materials within a common format. A brief account of the process of test development and the nature of the test materials will be given, together with discussion of the development of a scoring procedure using categories reflecting aspects of communicative performance. Results of analysis of data from two sessions of the roleplay-based subtest of speaking (N=200+ for each session) and one session of the subtest of writing (N=200+) using Item Response Theory will be presented. The analysis will present evidence for the validity of the assessment categories used in the scoring procedure.

In particular, discussion will focus on assessors' behaviour as revealed by the analysis and the extent to which candidates' scores reflect the orientation of assessors towards particular assessment criteria. The implications for the validity of communicative language tests will be considered. The analysis will demonstrate the use of IRT in investigating the validity of language tests as a complement to its more usual role in establishing test reliability.

Thom Hudson  
University of Hawaii at Manoa  

Testing the specificity of ESP reading skills

Reading programs and ESP reading projects in particular frequently focus on specific reading skills. However, skills and ability may develop differentially, both in terms of where they are located on a difficulty/ease continuum and in terms of how narrow or wide a spectrum of difficulty they encompass. That is, the ability to recognize correct tense/aspect from context may have a more narrow ability band than the ability to recognize logical connectors in an extended text or the ability to find main ideas in a reading passage. This differential band width has implications for such language testing concerns as the development of criterion-referenced tests and tests for placement of students in courses.

This study examines the test results for 500 native Spanish speaking chemical engineering students who took English for Science and Technology tests. It utilizes Rasch ability and difficulty estimates to examine the relationships among items which test: 1) the ability to determine grammar points from context; 2) reading comprehension of specific information and main ideas; and, 3) general reading ability as indicated on multiple-choice cloze tests. The test results are examined both in terms of the item difficulties of the three types of items and in terms of the degree to which examinee ability estimates differ depending upon which items are examined. The results indicate that certain skills, such as locating specific information, have a narrower range of difficulty than do other skills such as inferencing. Likewise, particular grammatical points, such as tense, are mastered at an earlier stage and thus have a narrower range than grammatical/discourse features such as logical connectors which are more globally determined. The discussion will focus on the implications such findings have for developing criterion-referenced tests aimed at specific domains and for placement exam development.
An ESL Proficiency Test for Teachers on Guam

On the Trust Territory of Guam, many teachers are nonnative speakers of English. However, English is the language of instruction at all levels. For several years, an ESL proficiency test has been given to teachers and aspiring teachers to determine their ability to instruct in English. Because this test enjoyed only limited success and acceptance, staff at the Center for Applied Linguistics were contracted to develop a new test that would serve as a valid measure of communicative competence in local K-12 educational settings. CAL staff worked in consort with staff at the University of Guam and the Guam Department of Education.

The Guamanian Educators' Test of English Proficiency (GETEP) measures language skills in English listening, speaking, reading, productive writing and editing. Each section is designed to assess the English language skills that would be utilized by teachers as part of their work. The listening, reading, and editing tests utilize a multiple-choice format. Editing skills are assessed through an error identification task utilizing a handwritten simulated student essay. The speaking test employs a direct interview format, and the productive, holistically-scored writing test assesses descriptive and persuasive writing.

Four parallel forms of the GETEP were administered to in-service and preservice educators in Guam in October, 1989. Following item analysis, the tests were revised. In order to validate the test and determine an appropriate passing score for each form, a committee of experts was designated by the Guam Department of Education. Using Angoff's Method 1 to determine a passing score, the judges carefully reviewed each item and indicated the probability that a minimally adequate teacher would answer the question correctly. These probabilities were summed and averaged across items and judges for each form of the test. The overall average probability became the passing score, or operational definition of minimally acceptable competence.

This paper will describe the development, validation, standard setting and implementation of the GETEP. Special attention will be given to a discussion of the standard setting procedure and its possible application in other situations.
Eva L. Baker, Center for the Study of Evaluation, UCLA
Jean L. Turner, TESL Department, UCLA
Frances A. Butler, Center for the Study of Evaluation, UCLA

An Initial Inquiry into the Use of Human Language Performance to Evaluate Artificial Intelligence Systems

In the spirit of this year’s testing colloquium theme, "collaboration and cooperation," this paper reports on research which brings together two areas of inquiry—language testing and artificial intelligence (AI). The major focus of the research was an attempt to reference intelligent computer programs, those that can process natural language, to the performance of humans. The goal was to begin to specify a continuum of difficulty for language an AI system could handle based on human test performances. To this end, natural language input which a specific AI system could process was analyzed linguistically and functionally. The analyses led to the development of the Natural Language Elementary Test (NLET) for pre-school and elementary school children.

The 126 subjects, 60 kindergarteners and 66 first graders, took two tests—the NLET and the language section of the Iowa Test of Basic Skills (ITBS), which was used as a criterion measure to allow for the grouping of students by national grade-equivalent norms. As expected, test results show steady improvement in NLET performance from the lowest to the highest grade equivalent band though there is considerable overlap in the range of NLET scores across grade-equivalent bands. The findings suggest that the NLET captures developmental linguistic skills in transition. Since NLET test specifications were developed from analyses of natural language input that was comprehensible to the AI system, an initial continuum of difficulty for the natural language understood by the AI system is proposed.
Lyle F. Bachman, UCLA (Coordinator)
Fred Davidson, Illinois State Board of Educ./Illinois Resource Center
John Foulkes, University of Cambridge Local Examinations Syndicate
Respondents:
   J. Charles Alderson, University of Lancaster
   John L.D. Clark, Defense Language Institute
   Bernard Spolsky, Bar-Ilan University
   Charles W. Stansfield, Center for Applied Linguistics

The Cambridge-TOEFL Comparability Study

The Cambridge-TOEFL Comparability Study (CTCS) investigated the comparability of two of the world's most widely-used EFL proficiency test batteries: the Test of English as a Foreign Language, Test of Spoken English and Test of Written English, administered by Educational Testing Service, and the First Certificate in English administered by the University of Cambridge Local Examinations Syndicate. Approximately 1,600 subjects from eight countries took both test batteries in December of 1989. Although the reliabilities of some of the tests were low, factor analyses of test scores indicate similar patterns of loadings across the two batteries. All measures loaded most heavily on a higher-order general factor, while factor loadings on the primary factors suggest that the most salient abilities measured by the two test batteries are speaking, listening, and a combination of reading, writing and structure. The factor loadings also suggest the presence of test method effects associated with each of the two test batteries.

The presentation of findings will be followed by critiques by recognized researchers in the field of language testing. The panel will discuss the implications of the findings for the interpretation of test scores from the two test batteries and for a theory of factors that affect performance on language tests. The viability of the CTCS as a model for future research will also be discussed.
Construct Validation of a Reading Comprehension Test:  
Combining Sources of Data

The purpose of this panel discussion is to present the results of a research effort that combined three methods of test analysis in an attempt to examine the construct validity of a reading comprehension test. The approach to test validation is broadened to include information from more sources. Item responses from 28 native speakers of Spanish studying at an intensive English program in the United States on the Descriptive Test of Language Skills -- Reading Comprehension subtest comprise the data. Three methods were used in the analysis. 1) Item responses from the 28 participants were submitted to traditional test performance analyses. The level of item difficulty as well as item discrimination were analyzed.

2) A content analysis of each test item on both forms of the Descriptive Test of Language Skills -- Reading Comprehension subtest was completed by applying the following four tests: (a) Bart and Read's (1984) statistical test to determine the prerequisite relationships that exist among the test questions; (b) Pearson and Johnson's (1978) question and answer relationship to determine if the relationship between the question and the text was textually explicit, textually implicit or scriptally implicit; (c) Schlesinger and Weiser's (1970) facet design theory to determine whether the test item required information from the text or from the reader's background knowledge; and (d) Meyer's (1975) prose analysis system to discuss how the items relate to the structure of the text.

3) Introspections from the 28 participants were analyzed to evaluate the strategies the readers had used to answer the comprehension questions. In order to preserve the timing factor which is integral to any testing situation, each participant was told that they would have 30 minutes to take the test as outlined in the test instructions. They were told that after reading and answering the comprehension questions for each passage they were to stop. The exam time was then suspended and participants were asked to provide a think-aloud protocol describing the reading and testing strategies they had used while reading that passage from the test. The time was then restarted as the participant continued reading and taking the test. The entire test was administered in this manner until a total testing time of 30 minutes has elapsed. Participants were allowed to produce their introspections in their native language.
Panel members will address each of the three analyses described above and explain how each leads to an understanding of readers' performance on the test. The use of all three methods provides a fuller perspective of the construct validation of this reading comprehension test.

This research also provides insight into the readers' use of strategies and whether they were successful in applying a strategy based on how the item is related to textually and non-textually based information.
Prompt difficulty, task type and performance level in ESL direct writing assessment

As direct writing assessment becomes an increasingly important factor in decisions to admit or place students in college and university academic programs, it becomes equally important to ensure that all who are tested have equal opportunities to demonstrate "optimal writing performance." (Tedick '89, p2.) The question to which the writer must respond (commonly called the "prompt") is a key variable in direct writing tests, with great potential to either enhance or distort such opportunities. Research to date, in both L1 and L2 writing assessment, has produced conflicting positions on this. (Carlison et al '85, Carlson and Bridgeman '86, Hamp-Lyons in press, Reid '89, Spaan '89, Tedick '89)

We still have much to learn about the effect of the prompt on writing performance in both L1 and L2 direct writing assessment. Essay scorers, and language teachers who prepare students for writing tests, often claim to know not only that some questions are easier or harder, but also which these are. This study investigated "expert" judgments of prompt difficulty in order to discover whether such judgments could be used as a source of information at the item writing stage of direct writing test development.

In this study, 68 prompts from the Michigan English Language Assessment Battery (MELAB, a test of English language proficiency which includes a 30-minute impromptu essay), were assigned difficulty levels by two trained MELAB essay scorers and two ESL writing experts. These judgments were then compared with the prompt performance data from scores on 8,583 MELAB essays, and the relationship between prompt difficulty and essay score was examined. The same judges also categorized the 68 prompts into several task types, and the performance data was analyzed to discover the relationship between prompt difficulty, task type and essay scores. The results showed measurable effects of topic difficulty and prompt type on score level, and of prompt type on prompt difficulty.

The results of this study shed additional light on such important questions as whether writers should be allowed a choice of prompt, whether certain types of prompts are better than others in eliciting a range of writing performance, whether item writing for a writing test requires a specially trained team of item writers, whether key features of fair, effective prompts can be isolated and described, and the extent to which piloting and revision need to be a central part of the writing test development process.
Prompt difficulty, task type and performance level in ESL direct writing assessment

As direct writing assessment becomes an increasingly important factor in decisions to admit or place students in college and university academic programs, it becomes equally important to ensure that all who are tested have equal opportunities to demonstrate "optimal writing performance." (Tedick '89, p.2.) The question to which the writer must respond (commonly called the "prompt") is a key variable in direct writing tests, with great potential to either enhance or distort such opportunities. Research to date, in both L1 and L2 writing assessment, has produced conflicting positions on this. (Carlson et al '85, Carlson and Bridgeman '86, Hamp-Lyons in press, Reid '89, Spaan '89, Tedick '89)

We still have much to learn about the effect of the prompt on writing performance in both L1 and L2 direct writing assessment. Essay scorers, and language teachers who prepare students for writing tests, often claim to know not only that some questions are easier or harder, but also which these are. This study investigated "expert" judgments of prompt difficulty in order to discover whether such judgments could be used as a source of information at the item writing stage of direct writing test development.

In this study, 68 prompts from the Michigan English Language Assessment Battery (MELAB, a test of English language proficiency which includes a 30-minute impromptu essay), were assigned difficulty levels by two trained MELAB essay scorers and two ESL writing experts. These judgments were then compared with the prompt performance data from scores on 8,583 MELAB essays, and the relationship between prompt difficulty and essay score was examined. The same judges also categorized the 68 prompts into several task types, and the performance data was analyzed to discover the relationship between prompt difficulty, task type and essay scores. The results showed measurable effects of topic difficulty and prompt type on score level, and of prompt type on prompt difficulty.

The results of this study shed additional light from a different direction on such important questions as whether writers should be allowed a choice of prompt, whether certain types of prompts are better than others in eliciting a range of writing performance, whether item writing for a writing test requires a specially trained team of item writers, whether key features of fair, effective prompts can be isolated and described, and the extent to which piloting and revision need to be a central part of the writing test development process.
THE ROLE OF INSTRUCTIONS IN TESTING SUMMARIZING ABILITY

The main purpose of the study was to determine the effects of specific guidelines in the taking of tests of summarizing ability -- tests in which respondents read source texts and provide written summaries as a measure of their reading comprehension level as well as of their writing ability. Another issue under consideration in this study was the effect of a rigorous scoring key on interrater reliability.

The subjects for this study were 63 native-Hebrew-speaking students from a teacher training college in Israel, 26 from two high-proficiency English-foreign-language (EFL) classes and 37 from two intermediate EFL classes. Four raters assessed the students' summaries in the study -- the two rating the Hebrew summaries of the Hebrew texts both being native Hebrew speakers, while of the two rating the Hebrew summaries of the EFL texts, one was a native Hebrew speaker and the other an English speaker. Five texts were selected for the study, two in Hebrew and three in English. Two sets of instructions were developed. One version was "guided" with specific instructions on how to read the texts and how to write the summaries. The other version had the typical "minimal" instructions. The scoring keys for the texts were based on the summaries of nine Hebrew-speaking and nine English-speaking experts respectively. All 63 respondents summarized the first Hebrew text, 53 summarized the second Hebrew text, and on the average, slightly more than a third of the students wrote summaries for the EFL texts.

The findings from the study suggested that the particular set of guidelines developed in this case had a mixed effect upon the respondents. An analysis of summaries on an item-by-item basis revealed that the guided instructions appeared to be both helpful and detrimental. In some cases, they assisted respondents in finding the key elements to summarize, and in other cases they probably dissuaded the respondent from including details that in fact proved to be essential in the eyes of the experts upon whom the rating key was based. With respect to interrater consistency, it was found that the raters differed in their ratings of several of the more global, linking ideas and of ideas involving details in the EFL texts. It would seem that these are perhaps the ideas that lend themselves to the most controversy in rating, even when a precise key is provided.
J Charles Alderson  
University of Lancaster

Judgements in Language Testing

Language testing is an area of applied linguistics that combines the exercise of professional judgement about language, learning and the nature of the achievement of language learning, with empirical data about students' performances and, by inference, their abilities. This paper addresses the issue of the relationship between judgements and empirical data in language testing by reporting on three studies. The first study compares the judgements that experienced test writers and examination markers make about the difficulty of items and tests, and compares those judgements with the results of an administration of the items. The second study investigates judgements by language professionals of test content and the skills and abilities supposedly being tested by certain test items, and compares these judgements with the results of test administrations and the introspections of test-takers. The third study gathers judgements from language testers and teachers about standards of performance of a given population, in a so-called standard setting exercise aimed at determining grade boundaries for a public examination. It compares the results at item and test level with the descriptive statistics of the examination's performance and discusses how judgements can assist in the setting of pass-fail distinctions. The paper ends by discussing the value of professional judgements in determining test appropriacy, test content and criterial cut-offs.
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 approximately 500 students in the English Second Language (ESL) programme and 1000 students in the French Second Language (FLS) programme. 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 programme designed to bring beginning second language learners to an intermediate level of proficiency. The programme consists of four instructionally linked 3 credits courses of which only the two higher level courses are credits courses. Such a programme 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 related 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.
A Comparative Analysis of Misfitting Items

This paper will present a comparative analysis of indices which identify items that contribute to a lack of reliability in measurement. The analysis will be based on a data set elicited from subjects from four proficiency levels who sat for a grammar/structure test. The indices to be studied include item difficulty, point biserial correlation, Sato's Caution index modified for items (Harnisch and Linn, 1981), and the Rasch item fit validity measure (Henning, 1987).

The paper will briefly describe each of the indices and provide a comparative analysis of them. Special consideration will be given to an examination of the proficiency-level differences for the indices. The paper will illustrate a potentially important use of such indices by identifying the items that contribute most to high values for the different proficiency levels and by providing the basis for judgment regarding the (in)appropriateness of the item content for the different proficiency groups.

A part of the paper will be a report of the following regression analysis. The p-values for each item will be computed for each proficiency group as well as for the entire sample. A linear regression will be performed on the p-values from the total sample. Residual scores will be computed by subtracting the expected proportion correct from the observed proportion correct for each proficiency level. The items will be categorized in terms of their content to determine the reasons for large differences in the residuals. The mean of the residuals for each category will be standardized by dividing by the standard error of estimate.
Tibor Von Elek, University of Gothenburg
Mats Oscarson, University of Gothenburg
Fred Davidson, Illinois State Board of Education; Illinois Resource Center

Implementation of National Composition EFL tests in Sweden: Pilot Phase.

For some time, there has been concern throughout Sweden that existing national EFL tests do not adequately sample the full EFL language skill range. For that reason, the Language Teaching Research Unit of the Department of Education at the University of Gothenburg was contracted to develop an experimental composition test for two secondary school levels. This test was to accompany other types of national EFL tests. This paper reports inter-rater reliability of these tests as well as on data analyses related to the following issue: are the composition and non-composition tests comparable enough to issue a single score for all components? Preliminary pilot results indicate a great deal of overlap and tend to say yes. However, certain national issues imply further study.

In that regard, the paper will also discuss the national educational measurement policy in Sweden and how it impinges upon the resolution of this issue. Sweden has, for many years, had a policy of nationally mandated grade levels which are defined in terms of a normal distribution curve, and any new component to any existing national test must add to the precision of the aggregate score. Therefore, there is great concern about the effect of the composition tests on overall rankings; this concern will be detailed and will be further discussed with the audience.
Bernard Spolsky
Bar-Ilan University

TOEFL - The pre-history

"This is not the first conference ever called on language testing, nor will it be the last one." With these words, John B. Carroll opened his classic paper at the conference held in Washington in May 1961 to establish an omnibus battery to test the English proficiency of foreign students applying to American colleges and universities. A study of this meeting and its background permits one to start answering a number of intriguing questions. Why, for instance, given the state of the art in language testing in 1961, and the fact that the major paper at the conference called for "integrative testing" did the recommendations not call for direct testing of speaking and writing? To what extent were the decisions of the conference influenced by linguistic theory and to what extent by psychometric principles? Or were practical and institutional considerations more important than theoretical ones? Answers to these questions cast light on the wider issue of the nature of applied linguistics, and in particular on the relative influence of theoretical ideas and of the institutional, social and economic situation in which the ideas are to be realized.
What the autonomous language learner can teach us about assessment

Over the past decade, literature on language learning and teaching has focussed increasingly on learner-centered language teaching, self-directed language learning, and autonomous language learning. (Dickinson, 1987; Riley, 1985) There is a general consensus in the literature that language learning is enhanced if the learner takes initiative in the language learning process and that responsibility for the management of language learning does not fall exclusively on the language teacher. Other studies (Naiman, Frolich, Stern & Todesco, 1978; Rubin, 1975) have investigated what the good language learner can tell us about successful learning strategies in the management of language learning. Our research asks what the autonomous language learner can teach us about the role of assessment in language learning.

This study focuses on the approach that autonomous language learners use to assess the quality of their language learning program, the extent of their progress (the product of their learning), and the effectiveness of their learner strategies (the process of their learning). We are interested in the relationship between autonomy and control over assessment. Like Canale (1986), we wish to explore how a language learner can avoid becoming no more than "an obedient examinee, a disinterested consumer, a powerless patient, or even an unwilling victim" (p.250).

Our paper contains indepth profiles of four autonomous language learners, a reexamination of our understanding of the terms autonomy and self-direction using control over assessment as the distinguishing feature, a description of a model of the language learner in which control over assessment is pivotal, and an examination of the relevance of this model for decisions important to the management of assessment in the classroom. Because research of this nature is at an exploratory stage, we have chosen a qualitative rather than quantitative research design in order to explore fully what questions and issues might inform further research. We believe that such an analysis will be useful to educators. As Dickinson, (1987:1) notes in his discussion on effectiveness of self-instruction in language learning, "the results do not exist because the research has not been done." Any such research, we argue, must address the crucial relationship between assessment and learner autonomy.
PARTICIPANTS AND PAPER PRESENTERS - LTRC SAN FRANCISCO 1990

ORGANIZER: LIZ HAMP-LYONS,
U OF MICH, 1025 Angel Hall,
Ann Arbor, MI 48103

ABSTRACTS: DARIUSH HOOSHMAN
DLIFLC/ATFL-GST Presidio
Monterey, CA 93944-5006

LOCAL ARRANGEMENTS: DONNA ILYIN
1136 CLEMENT ST., S.F. CA. 94118

J.C. ALDERSOHN
DEPT OF LING & MOD ENG LANG
University of Lancaster
Lancaster LA1 4YT
UK

JUDGEMENTS IN LANG TEST;
& PANEL ON THE ELTS
(also see Bachman Panel
on Camb-TOEFL Study)

NEIL ANDERSON
DEPT OF LING
GORDY HALL OHIO U
ATHENS
OH
45701

CONSTRUCT VALIDATION OF A
READING COMPREHENSION
TEST COMBINING SOURCES OF
DATA

LYLE BACHMAN
U OF CALIF AT LA
3300 ROLFE HALL
LA
CA
90024

PANEL ON THE CAMBRIDGE-
TOEFL COMPARABILITY STUDY
(See also Anderson Paper)

J.D. BROWN
Department of ESL
University of Hawaii
Honolulu
HI
96822

A COMPREHENSIVE CRITERION-
REFERENCED LANGUAGE TESTING
PROJECT

FRANCES A BUTLER
2517 Gates Ave., Unit A
Redondo Beach
CA
92078

(See Jean Turner)

EDUARDO C. CASCALLAR
ETS 10-P
PRINCETON
NJ
08541

THE ASSESSMENT OF LANGUAGE
APTITUDE AND THE PREDICTION
OF ACHIEVEMENT: DEVELOPMENT
OF A COMPONENTIAL APPROACH
TO THEORY AND TESTING

CAROLINE CLAPHAM
LANCASTER UNIV
BAILRIGG
LANCASTER
UK
LA1 4YT

See Alderson: Panel on ELTS
JOHN L.D. CLARK  
DLI, Foreign Language Center  
ATFL-DES-T  
Presidio of Monterey  
CA.  
93944

See Eduardo Cascaller; also see Bachman: Panel on Camb-TOEFL study.

ANDREW COHEN  
School of Education  
Hebrew University  
91905 Jerusalem  
Israel

THE ROLE OF INSTRUCTIONS IN TESTING SUMMARIZING ABILITY;
Also see Neil Anderson

FRED DAVIDSON  
15BE/IRC  
1855 MT PROSPECT RD  
DES PLAINS  
IL  
60018

See Bachman: Panel on Camb.-TOEFL study; Also see Oscarson

JOHN H.A.L. DE JONG  
CITO NAT INST FOR ED MSMT  
P.O. Box 1034  
6801 Mg Arnhem  
The Netherlands

TEST DIMENSIONALITY IN RELATION TO STUDENT PROFICIENCY

DAN DOUGLAS  
DEPT OF ENG. 203 ROSS HALL  
Iowa State University  
Ames  
IA  
50011

PERFORMANCE ON GENERAL VERSUS FIELD-SPECIFIC TESTS OF SPEAKING PROFICIENCY

JOHN FOULKES  
UCLES  
1 Hills Road  
Cambridge  
England  
CBI 2EU

See Alderson: Panel on ELTS

GORDON HALE  
Educational Testing Service  
Princeton  
NJ  
08541

NOTE TAKING AND TOEFL LISTENING COMPREHENSION

GRANT HENNING  
RESEARCH DIV 10-P  
ETS  
Princeton  
NJ  
08541

A COMPARISON OF MULTIPLE-CHOICE AND OPEN-ENDED COMPUTERIZED ASSESSMENT OF ESL READING COMPREHENSION
THOM HUDSON
U OF HAWAII AT MANOA
1890 EAST-WEST ROAD
HONOLULU
HI
96822

DORRY KENYON
CAL
1118 22ND STREET
WASH.
D.C.
20037

TIM MCNAMARA
U OF MELBOURNE
PARKVILLE
VICTORIA 3052
AUSTRALIA

MATS OSCARSON
Gothenberg University
BOX 1010
Molndal
SWEDEN
5-43126

BONNY NORTON PEIRCE
OISE
308 WENDY LANE
OALWILLE, ONT
CANADA
L6L 4RI

KYLE PERKINS
Linguistics Department
Southern Illinois University
Carbondale
IL
62901

SHEILA PROCHNOW
U OF MICH
ELI 3001 N U BLDG
ANN ARBOR
MI

MARY LEE SCOTT
CAL
1118 27TH ST. NW
WASHINGTON
DC
20037

TESTING THE SPECIFICITY
OF ESP READING SKILLS

See Stansfield: Guam
Proficiency Test

ITEM RESPONSE THEORY AND
THE VALIDATION OF AN ESP
TEST FOR HEALTH
PROFESSIONALS

IMPLEMENTATION OF NATIONAL
COMPOSITION ESL TESTS IN
SWEDEN: PILOT PHASE

WHAT THE AUTONOMOUS
LANGUAGE LEARNER CAN TELL
US ABOUT ASSESSMENT

A COMPARATIVE ANALYSIS OF
MISFITTING ITEMS

Also see Anderson

PROMPT DIFFICULTY, TASK
TYPE AND PERFORMANCE
LEVEL IN ESL DIRECT
WRITING ASSESSMENT

See Stansfield: Development
and Validation of a
Personnel Selection Test of
Spanish Listening and
Summary Writing
LARRY SELINKER
English Language Institute
University of Michigan
Ann Arbor
MI
48109

ELANA SHOHAMY
School of Education
Tel Aviv University
Ramat Aviv
Israel

MARY SPAAN
ELI-U OF MICH
6085 TRIST ROAD
GRASS LAKE
MI
49240

BERNARD SPOLSKY
ENGLISH
Bar Ilan University
RAMAT-GAN
Israel

CHARLES W. STANSFIELD
Center for Applied Linguistics
1118 22nd Street, N.W.
Washington
D.C.
20037

JEAN TURNER
UCLA
2116 OAK ST., APT C
SANTA MONICA
CA
90405

MARIJKE WALKER
FBI HQ, Room 3505
10TH & PENN AVE NW
Washington
DC
20535

See Douglas

A DIAGNOSTIC FEEDBACK MODEL
FOR ASSESSING ACHIEVEMENT
AND PROFICIENCY OF FOREIGN
LANGUAGES

THE EFFECT OF THE PROMPT
IN ESSAY EXAMS

TOEFL HISTORY
See also Bachman:
Panel on Camb-TOEFL
Study

DEVELOPMENT AND VALIDATION
OF A PERSONNEL SELECTION
TEST OF SPANISH LISTENING
AND SUMMARY WRITING; &

AN ESL PROFICIENCY TEST FOR
TEACHERS ON GUAM
Also see Cascallar and
See Bachman: Panel on
Camb.-TOEFL study;

AN INITIAL INQUIRY INTO THE
USE OF HUMAN LANGUAGE
PERFORMANCE TO EVALUATE
ARTIFICIAL LANGUAGE
See Stansfield: Development
and Validation of a
Personnel Selection Test of
Spanish Listening and
Summary Writing