Fourteenth Annual
Language Testing Research
Colloquium

February 27 to March 1, 1992
Asian Studies Centre
University of British Columbia
Vancouver, British Columbia
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conference Schedule</td>
<td>1</td>
</tr>
<tr>
<td>Abstracts for the Paper Sessions</td>
<td>7</td>
</tr>
<tr>
<td>arranged alphabetically</td>
<td></td>
</tr>
<tr>
<td>Abstracts for the Poster Session</td>
<td>19</td>
</tr>
<tr>
<td>arranged alphabetically</td>
<td></td>
</tr>
<tr>
<td>Abstracts for Software Applications/Demonstrations</td>
<td>25</td>
</tr>
<tr>
<td>arranged alphabetically</td>
<td></td>
</tr>
<tr>
<td>Symposia: Overview and Abstracts</td>
<td>27</td>
</tr>
<tr>
<td>in the order of presentation</td>
<td></td>
</tr>
<tr>
<td>Addresses of Program Participants</td>
<td>31</td>
</tr>
</tbody>
</table>
Special Acknowledgements

1992 Colloquium Organizing Committee
John de Jong, Grant Henning and Marjorie Wesche (Chair)

Local Arrangements
Rick Berwick, Alister Cumming and Helen Tegenfeldt

LTRC Archives
Fred Davidson

Abstract Readers
Charles Alderson
Lyle Bachman
Caroline Clapham
Fred Davidson
John de Jong
Dan Douglas
Liz Hamp-Lyons
Grant Henning
Thom Hudson
Michel Laurier
Bonnie Norton-Pierce
Doreen Ready
Elana Shohamy
Charles Stansfield
John Upshur
Marjorie Wesche

Program Booklet
Doreen Ready and Marjorie Wesche

Preparation of Documents and Mailing List Maintenance
Beatrice Magyar

Registration
Helen Tegenfeldt, Beryl Tonkin and Marjorie Wesche

Wine and Cheese Reception and Colloquium Refreshments
University of British Columbia, Language Education Department

Buses to Vancouver Harbour
Vancouver Community College

Administrative Support
Second Language Institute, University of Ottawa
Language Education Department, University of British Columbia
Vancouver Community College

Publication of Selected Papers
Rick Berwick and Alister Cumming

And a special thanks to all presenters, chairpersons and all volunteers
4:00 - 6:00 p.m.  Registration

7:30 - 9:30 p.m.  Wine & Cheese Reception
(Hosted by UBC Language Education Department)

FRIDAY, FEBRUARY 28

7:30 - 8:30 a.m.  Registration coffee, juice and muffins

8:30 - 8:45 a.m.  Opening and Introduction (M. Wesche)

8:45 - 9:30 a.m.  Plenary: "Task-Based Testing" P. Skehan

PAUSE

9:45 - 12:00 noon  Paper Session I. Chair: J. de Jong

C. Clapham: "What Makes an ESP Reading Test Appropriate for its Candidates?"

C. Elder: "How do Subject-Specialists Construe Second Language Proficiency?"

J. Hulstijn: "Individual Differences in L2 Proficiency as a Function of Individual Differences in L1 Proficiency"

K. Perkins, G. Henning: "Differential Person Functioning on Language Proficiency Tests"

12:00 - 1:00 p.m.  LUNCH (buffet tickets available)*

1:00 - 2:15 p.m.  Paper Session II. Chair: B. Spolsky

C. Stansfield, D. Kenyon: "Comparing the Scaling of Speaking Tasks by Language Teachers and by the ACTFL Guidelines"

C. Taylor, G. DeMauro, B. Voltmer, B. Henderson: "The Impact of TWE Reader Qualifying Criteria on Reader Performance"

PAUSE

2:30 - 4:15 p.m.  Symposium: "Educational and Social Impact of Language Tests".

Chair: B. Spolsky

E. Shohamy**, E. Cascallar**, C. Alderson, D. Wall. Discussants

G. Henning, B. Spolsky.

PAUSE

4:30 - 6:00 p.m.  Test Software Applications/Demonstrations

Chair: G. Henning

M. Piemann, I. Thornton, A. Mackey: "On-line Diagnostic Screening of Interlanguage Acquisition by Expert System"

M. Des Brisay: "Applications of TESTGRAF in Setting Cut-off Points on ESL Tests"

M. Laurier: "Using the Information Curve to Assess Language CAT Efficiency"

* Lunch tickets must be purchased at registration
** Symposium organizers
8:00 a.m.  Coffee, juice, muffins

8:30 - 10:15 a.m.  Paper Session III. Chair: D. Porter
                      A. Brown: "The Role of Test-Taker Feedback in the Test Development Process"
                      C. Gordon: "The Effects of Affective Variables and Testing Methods on Test Performance in EFL"
                      I. Wijg: "A Communicative Test in Analysis"

                      PAUSE

10:30 - 12:15 a.m.  Paper Session IV. Chair: L. Hamp-Lyons
                      S. Ross: "Accommodative Questions in Oral Proficiency Interviews"
                      E. Shohamy: "Discourse Validation of a Direct Versus a Semi-Direct Oral Test"
                      R. Young, M. Milanovic: "Discourse Variation in Oral Proficiency Interviews"

                      LUNCH (box lunches available)*

12:30 - 1:45 p.m.  LTRC Business Meeting "Brown Bag"

1:45 - 2:00 p.m.  Report on International Language Testing Association (C. Stansfield)

2:00 - 2:15 p.m.  Presentation on Language Testing (A. Davies, J. Upshur, eds.)

2:30 - 4:45 p.m.  Paper Session V. Chair: J. Upshur
                      G. Buck: "The Analysis of Multidimensional Data Sets"
                      J. de Jong: "Methodology for Formally Combining Dichotomously and Polychotomously Scored Items"
                      G. Henning: "Dimensionality and Construct Validity of Language Tests"
                      T. McNamara, A. Brown, C. Elder, T. Lumley: "Mapping Abilities and Skill Levels Using Rasch Techniques"

5:30 p.m.  Buses to Vancouver Harbour from Gage Court and Asian Studies Centre

3 - 4 hours  Banquet and Harbour Cruise

* lunch tickets must be purchased at registration
8:00 a.m. Coffee, juice and muffins

8:30 - 10:45 a.m. Paper Session VI. Chair: A. Davies
J.D. Brown, E. Detmer, T. Hudson: "Developing and Validating Tests of Cross-Cultural Pragmatics"
J. Fox, B. Zumbo, T. Pychyl: "Psychometric Properties of the CAEL Assessment: An Examination of the Dependability/Reliability of Placement Decisions"
D. Lussier: "A Systemic Approach to Evaluation of Second Language Programs in Quebec"
C. Turner, J. Upshur: "Validation of Measures of Grammatical Knowledge and Communicative Ability"

PAUSE

11:00 - 12:15 p.m. Poster Presentations. Chair: G. Buck
H. Tegenfeldt, V. Monk: "The A-LINC Assessment"
M. Wesche, D. Ready: "The Ontario Test of ESL: OTESL"
M. Des Brisay, J. St.John: "Canadian Test of English for Scholars and Trainees: CanTEST"
J. Ross, S. Chyn: "TOEFL 2000: Future Directions"
E. Cascallar, M. Walker: "TSE Validation in Federal Government Language Programs"
S. Cushing, B. Lynch: "Hypothesis Testing in Construct Validation"
A. Cumming, D. Mellow: "Validity of Written Indicators of Second Language Proficiency"
D. Douglas, L. Selinker: "CHEMSPEAK: An Update"
A. Huhta, E. Randell: "Multiple-Choice Summary: A Measure of Test Comprehension"
J. Hulstijn: "Implementation Issues: National Dutch Language Certification Examinations"

12:15 - 2:00 p.m. LUNCH and Poster displays (box lunches available)*

2:00 - 3:15 Paper Session VII. Chair: E. Shohamy
K. Bailey, J.D. Brown: "Language Testing Courses: What are They?"
A. Davies: "The Role of the Segmental Dictionary in Professional Validation"

3:30 - 5:15 p.m. Symposium: "Development and Use of Rating Scales in Language Testing". Chair: G. Henning
D. Kenyon**, M. Milanovic, N. Saville, A. Cook, T. McNamara, J. Hamilton, E. Sheridan, F. Davidson, L. Hamp-Lyons, B. Tyndall

5:15 - 5:30 p.m. ANNOUNCEMENTS

5:30 p.m. CLOSURE
* lunch tickets must be purchased at registration
** symposium organizer
Papers

Language Testing Courses: What are They?

Kathleen Bailey, James Brown

There is an increasing trend for language testing courses to be offered as part of the curriculum in the ESL teacher training programs around the world. The purpose of this project was to investigate the structure, content and student attitudes toward introductory language testing courses. To those ends, a questionnaire was designed to cover a variety of topics including the instructor’s background, the topics covered in the course, the types of students in the classes, as well as the students’ attitudes toward language testing before and after the course. While the questions were predominantly Likert-scale in format, a number of open-ended questions were also included.

The questionnaire was sent to all the "active members" on the mailing list for the Language Testing Research Colloquium during Fall semester 1990. Two months later, a second mailing was sent out to those active members who had not responded to the first request for information.

There were a total of 150 respondents. Of these, more than half (n = 76) indicated that they had never taught a testing course (or had not taught one for a number of years). The remaining 74 respondents indicated that they had taught such a course; these 74 all completed the questionnaire. The results indicate considerable variation in the training of the instructors as well as in the shape that such testing courses take. The data are described in terms of what the typical testing course looks like, and are then cross-tabulated to show differences and similarities in teacher background, content, and student attitudes among the various types of institutions.

The Role of Test-Taker Feedback in the Test Development Process

Anne Brown

Recent research into performance test development (Kenyon and Stansfield 1991; Alderson 1988) raises the issue of the role of test-taker feedback in the test development process.

The value of such feedback is investigated in the context of a project to develop an LSP oral/aural proficiency test in Japanese for Tourism and Hospitality. Fifty three subjects drawn from a range of Japanese language courses undertook the trial version of the test and completed a post-test questionnaire in which they provided reactions to the test as a whole as well as to task types and individual test items. Reactions consisted of a mixture of open comments and ratings on a 5-point scale.

Analysis of the data delineated the relationships between the characteristics of test-takers (such as relevant occupational experience, amount of language study completed, familiarity with the test format, etc.) and their reactions to the test.

Examination of the comments provided by the test-takers on individual items together with the statistical analysis of the test proved to be of value in the item revision process. Other comments enabled substantial improvements to be made to the overall design of the test and the test handbook.

The paper concludes with a discussion of the implications of the research findings for test validity.
Developing and Validating Tests of Cross-Cultural Pragmatics

J.D. Brown, Emily Detmer, Thom Hudson

There are numerous existing methods for measuring grammatical and textual competence, but no generally accepted valid measures of communicative ability components such as pragmatic competence. This paper presents results of a project to develop and validate tests of cross-cultural pragmatic ability.

The design is a multitrait-multimethod study using the traits of power, social distance and imposition in the speech acts of requests, refusals and apologies. Three different classifications of test methods are used: indirect measures of pragmatic knowledge (open ended and multiple-choice), direct measures of pragmatic ability (role play and structured interview), and self-assessment of pragmatic skill (self-rating and evaluation of performance on video).

Subjects of this study are native speakers of Japanese in the U.S. and native speakers of English in Japan. In Spring 1991, pilot versions were administered in the U.S. and are being analyzed. In Fall 1991, students in Japan and the U.S. will be administered the indirect paper and pencil tests. From those results a smaller sample will be selected for further testing using the simulation and oral interview. All students who take the paper and pencil tests will be administered the cued response self-assessment instruments. The results of the tests will be analyzed: 1) through a content analysis comparing native with non-native speaker responses; 2) statistically with a focus on the effectiveness of individual items; 3) for trait validity and method effect.

Implications for validity testing cross-cultural pragmatic ability will be discussed.

The Analysis of Multidimensional Data Sets

Gary Buck

Most of the statistical procedures used in language testing require the assumption that the data is unidimensional. This assumption underlies the use of both IRT models, and classical procedures such as internal consistency estimates and item/total correlations. Indeed, adding together item scores is only meaningful if like is being added to like. However, theoretical and empirical work in cognition, linguistics and language testing suggests overwhelmingly that language performance depends on a complex interaction of a multitude of variables, indeed anything within human experience is a potential influence, and hence each individual item response in language test data is a unique, multidimensional composite of numerous variables.

Clearly statistics used in language testing are robust to some violation of their assumptions, but the question of importance is whether the multidimensional nature of language test data violates these assumptions sufficiently to invalidate their use. This is an empirical question, which is addressed in this paper by the analysis of a number of error-free, 60-person-by-60-item data sets, created to progressively and systematically violate the unidimensional assumption. Using a model of multidimensional data from Coombs (1964), dimensionality will be operationalized within items, not between items. Bi-dimensional, tri-dimensional and multidimensional data sets were created with traits combined in conjunctive, disjunctive, compensatory and mixed composition. The 18 data sets will be analyzed using classical test analysis and procedures based on IRT models. They will also be subjected to the more common tests of dimensionality. Results will be presented, and implications for language testing will be discussed.
What Makes an ESP Reading Test Appropriate for its Candidates?

Caroline Clapham

Researchers into the effect of background knowledge on the reading test performance of university students have tended to presume that a reading text is specific for students in a given subject area if it is based on a topic within that subject area. However, this is not necessarily the case. In a study of student reading performance in three different academic fields, Clapham (1990) showed that not all the supposedly subject-specific reading passages were as specific as intended. Some were appropriate for students in more than one of the three designated subject areas, one was inappropriate for the targeted students, and one appeared to be more appropriate for students outside the relevant academic area. In an attempt to find out why this might be, a group of judges (some were language testers, some academic subject specialists) used a modification of Bachman’s Test Method Characteristics and Communicative Language Ability rating instruments (Bachman 1991) to categorise the content of 10 nominally subject-specific reading tests. Texts and tasks were analyzed according to the criteria of source text, subject area, topic, genre, propositional content, degree of contextualisation, and organisational and grammatical complexity. The results were compared to measures of the tests’ appropriacy which had been gained from students’ test scores and questionnaire responses, and the effect of each of the criteria on test appropriacy was considered.

This paper reports on these findings and uses them to draw up proposals for the practical selection of suitable passages for ESP reading tests.

The Role of the Segmental Dictionary in Professional Validation

Alan Davies

The paper argues that a major aspect of language test validation is the deliberate construction of a profession of language testers through education and training, including the formation of an association, the establishment of journals etc., thereby fostering agreement on goals, procedures, terminology, norms and methods of evaluating innovations. An important tool of professionalising is the writing of a specialist dictionary (Crystal 1985; Richards, Platt and Weber 1985), also called a segmental dictionary (Opitz 1983), fulfilling a similar role to dictionaries in language standardisation.

A collaborative project in writing a language testing dictionary is described, the aim being to “try and introduce a measure of normalisation in the use of specialist terms and thus facilitate the exchange of information” (Moulin 1983, 146).

Distinctions between a dictionary, an encyclopedia and a glossary are considered. Feedback suggests that what is needed is an “encyclopedic dictionary” (McArthur 1986) since the project defines and avoids essays (and so is not a glossary, Abrams 1981) but has no information on pronunciation or on historical derivation.

Problems of coverage and of entry style are discussed and readers’ feedback on sample entries compared. Further offers to trail and critical evaluation of entries by LTRC participants are invited.
Methodology for Formally Combining Dichotomously and Polychotomously Scored Items

John de Jong

The method is illustrated with an empirical example based on data on a semi-direct oral proficiency test. The test consisted of 5 subtests, but only two of these will be used for the illustration: the first containing 20 dichotomously scored items on pronunciation, the second consisting of 10 items on a simple dialogue task where answers are rated in three ordered categories.

While the mathematical extension of the simple logistic model (the Rasch model) to the extended logistic model is quite straightforward, there is a potential cost as well as a benefit in incorporating polychotomously scored items. The potential benefit is that for the same stimulus, a more complex response and a finer discrimination can be obtained: the potential cost, and one that can be disclosed by the methodology, is that the ordering of the categories is not working as intended.

In the analysis it is shown that for some items the intended ordering of the categories was not as expected. Post hoc explanations of this malfunctioning are offered as well as suggestions for item revision resulting from the analysis. Results of a second administration of the test after revision will permit an evaluation of the effect of the revisions and therefore also of the appropriateness of the interpretation of the analyses from the first administration.

How do Subject-Specialists Construe Second Language Proficiency?

Catherine Elder

Recent research on rater variance (e.g., Barnwell 1990; Douglas and Selinker 1990; Hadden 1991) raises the question of whether "linguistically naive" subject-specialists may be better equipped than language experts to judge the effectiveness of particular areas of non-native speaker communication.

This question is investigated in the context of a project on the development of a classroom-based observation schedule to assess the English language proficiency of non-native speaker graduates training as secondary mathematics and science teachers. The paper examines aspects of rater behaviour as evidenced in recent trials.

The schedule was applied to observations of actual performance in the maths and science classroom as well as to the viewing of a number of videoed segments of classroom interaction. Ratings were elicited from two groups of assessors: 12 ESL teachers and 8 subject specialists (maths/science teachers/teacher-trainers). This allowed estimation of the intra- and inter-group reliability of the procedure and ultimately of the validity of using non-language experts as judges of language proficiency.

While there are significant correlations between subject specialists' and language teachers' overall judgements of communicative effectiveness, the application of t-tests and factor analyses to the data reveal differences between the two groups with respect to their ratings of particular dimensions of language use and the weighting of these dimensions in relation to global proficiency assessments. The paper concludes with a discussion of the theoretical and practical implications of these findings as they relate to the larger issue of the validity and reliability of occupation-specific performance tests.
Psychometric Properties of the CAEL Assessment:  
An Examination of the Dependability/Reliability of Placement Decisions

Janna Fox, Bruno Zumbo, Tim Pychyl

In this paper we will explore some of the psychometric properties of the Carleton Academic English Language (CAEL) Assessment. As this criterion-referenced assessment is used to make placement decisions (including eligibility for university admission) for non-native speakers of English at Carleton University, key issues are the reliability and validity of these decisions.

Using H. Huynh's domain referenced reliability methods, we will examine the reliability of the placements made based on CAEL Assessments between 1988-1991 (n = 1000). Data based on the qualitative assessment of student placement by EAP teachers and students' course performance will also be presented. The results will be discussed in relation to the validity of this type of criterion-referenced test for making placement decisions in EAP programs.

The Effects of Affective Variables and Testing Methods on Test Performance in EFL

Claire Gordon

The purpose of this study is to investigate the relationships between affective variables — attribution of success and failure in EFL tests, test anxiety and attitudes toward learning English as a foreign language — and performance on multiple choice and open ended tests in EFL reading comprehension tests. The literature on the attributional theory of motivation suggests that causal ascriptions of success and failure in school performance give rise to affects and future expectancies, and these in turn are related to achievement outcomes. Research on the relationship between test anxiety and performance on cognitive tasks has produced inconsistent results with reports of both beneficial and detrimental effects of anxiety on achievement. In general, studies which have investigated the relationship of these psychological factors to school performance have focused mainly on the area of mathematics. The present study aims to test the theory for performance in EFL reading comprehension tests with particular attention to possibilities of different patterns of relationships for different test formats — multiple choice (MC) and open ended (OE).

Research Questions
1) Does receiving the test form of one's preference affect test performance?
2) Do students of different ability levels prefer different test formats or perform differently on different test forms?
3) How are affective variables related to performance on the two different test forms?

Subjects
The sample consists of 150 university students studying reading comprehension in EFL courses at the intermediate level.

Instruments
1) Two forms of a reading comprehension test (one open ended and one multiple choice) were administered randomly. Each test contained the same three short reading texts with a total of 14 items.
2) A questionnaire measuring attribution and attitude variables was administered prior to the test.
3) The Hebrew version of the Sarason RTT test anxiety questionnaire was administered prior to the test.
Dimensionality and Construct Validity of Language Tests

Grant Henning

Methods for establishing test dimensionality have received considerable research study in recent years (e.g., Boldt 1989; Goldstein 1980; Hambleton and Swaminathan 1985; Hattie 1985; Henning 1988; Oltman and Stricker 1990). This attention is appropriate given that the assumption of unidimensionality underlies the application of item response theory (IRT) and classical true score theory to test construction, test equating, and validity of ability inferences drawn from test scores. However, some researchers have lamented the constraints imposed by unidimensionality assumptions on language test content variation (e.g., Bachman 1989). Still others have asserted that IRT methodology with its unidimensionality assumption may be inappropriate for the study of complex multidimensional domains such as communicative language ability.

The present study will offer simulation evidence that "psychometric" dimensionality as measured by a variety of techniques may be somewhat independent of the number of psychological domains or constructs measured. An analysis of 10 simulated data sets, each involving 30 items with 100 examinees, will be presented to demonstrate that psychometric unidimensionality assumptions may be satisfied in situations where multidimensional constructs are measured simultaneously by the same test. It will also be shown that some techniques for testing psychometric dimensionality may suggest psychometric multidimensionality when there is clearly only one construct or dimension measured by the test. The conclusion is offered that psychometric dimensionality is sample dependent and that use of IRT methodology may not be inappropriate for tests of communicative language ability even though complex underlying constructs underlie performance on the test.

Individual Differences in L2 Proficiency
as a Function of Individual Differences in L1 Proficiency

Jan Hulstijn

In this presentation I will argue that individual differences in L2 proficiency should be conceived of as a partial function of individual differences in L1 proficiency, and that the relationship between L2 and L1 proficiency has for too long been neglected by L2 learning researchers.

To illustrate this claim, I will give a brief review of the relevant empirical literature. In particular, I will describe two studies which were conducted at our university. The first study dealt with hesitation phenomena in L2 and L1 speech. Four groups of Dutch learners of English (n = 65) read an English (L2) and a Dutch (L1) text aloud. The independent variables were English proficiency (high/low), and Grade (9/11). The dependent variables were Reading Speed, Prediction of 12 words omitted from the text, Repeats, Self Corrections and Deviations from the text.

The second study dealt with reading comprehension in L2 and L1. Subjects in this study were 60 adult, educated Turkish learners of Dutch, living in the Netherlands. They performed reading comprehension tests in Dutch (L2) and Turkish (L1), as well as a L2 vocabulary test and a L2 grammar test. Regression analyses were conducted to predict L2 reading comprehension performance on the basis of L1 reading comprehension, L2 vocabulary knowledge and L2 grammar knowledge.

The results of these studies suggest that differences between L2 learners on L2 tasks should be adjusted for differences on corresponding L1 tasks, before drawing conclusions pertaining to second language proficiency as a theoretical construct. Such a perspective could be useful for language teachers and language testers alike, highlighting the role of non-L2-specific factors in L2 proficiency.
A Systemic Approach to Evaluation of Second Language Programs in Quebec

Denise Lussier

This presentation will be an overview of a systemic approach to second language program evaluation. This approach was seen as an integral part of a curriculum revision process. It was initiated by the Quebec Ministry of Education to make informed decisions with respect to retaining or revising the program in French as a Second Language. This project started in April 1989 and involved all those concerned with the teaching of FSL (administrators, curriculum advisors, parents, teachers and students).

The development and implementation of instruments for the project will be discussed. The focus is on the second language criterion-referenced test administered to measure the learning outcomes based on program objectives. The subjects were 600 junior high school students.

Reliability and validation measures will be discussed taking into consideration the parameters identified in the Definition of Domain of the instrument. Discussion will also include procedures for data collection concerning language skill mastery levels, identification of performance standards in a communicative-based program, and the establishment of a threshold level.

As such, this systemic process sought to assess the validity of the program learning objectives. It also investigated the program effects on teaching conditions and practices. Finally, it identified to what extent the students were able to meet the learning outcomes of the study program.

Mapping Abilities and Skill Levels Using Rasch Techniques

Tim McNamara, Anne Brown, Catherine Elder, Tom Lumley

This paper considers the use of the results of Rasch IRT analysis of data from tests of reading and listening skills in certification, diagnosis and placement.

A central concept in Rasch IRT analysis is the mapping of item difficulty and person ability onto a single scale. This allows the possibility of empirical definition and validation of levels of achievement or proficiency in the skill being assessed, represented by the clustering of items testing similar sub-skills at points along the ability/difficulty continuum defined by the analysis. It further allows description of the ability of individuals or groups in terms of the skills so identified. The mapping process can thus be used to generate criterion statements of skill levels for purposes of certification, selection or placement. Individual Student Profiles can also be generated to provide diagnostic information on individual candidates. The potential of Rasch analysis in this regard is illustrated with examples from two recent test development projects: one a scheme to promote the learning of foreign languages in the junior secondary school, the other a test of English for academic purposes.

Crucial to the above uses of this mapping procedure is the association of sets of items of given difficulty with specific kinds of sub-skills. How straightforward or problematic is this? Research on the relationship of item types to difficulty levels on a 55 item subtest of reading comprehension in a new EAP test is reported, and the difficulties encountered discussed. A proposal for a program of further research in this area is defined.
Differential Person Functioning on Language Proficiency Tests

Kyle Perkins, Grant Henning

Considerable attention has been given to recent studies of differential item functioning (DIF) on language proficiency tests (e.g., Alderman and Holland 1981; Chen and Henning 1985; Kunan 1990). These and other studies employ a variety of methods to examine whether persons of matched ability drawn from different language, gender, or ethnic groups perform differently on the same test items. In this way, such studies seek to identify systematic item performance differences that may inform the study of person groups. The presence and direction of such item performance differences may also impact upon the validity of inferences drawn about person ability on the basis of test scores.

Less attention has been given to the related phenomenon of differential person functioning (DPF) on language proficiency tests (e.g., Henning 1990). DPF studies imply DIF methodology; but unlike DIF studies, DPF studies ask whether items of matched difficulty drawn from different format or content domains function differently for the same persons. The presence of systematic DPF may signify test multidimensionality and may have implications regarding construct validity.

The proposed study will examine differential person functioning for ESL/EFL students responding to TOEFL test items of matched difficulty drawn from different content/format domains of the TOEFL test. Suggestions will be offered about alternative DPF methodologies and about item difficulty regions and content domains where DPF is most evident. This study is also intended to contribute to a better understanding of domain theory.

Accommodative Questions in Oral Proficiency Interviews

Steven Ross

The processes involved in defining oral proficiency in second languages have to date not involved detailed analysis of the discourse characteristic of oral proficiency interviews. The present study considers the phenomenon of variation in questions posed by interviewers at key junctures in the interview process. Based on variable rule analyses of sixteen full-length oral proficiency interviews, it is argued that perceptions of oral proficiency are reflected in the extent of accommodation in interviewer questioning, and that the extent of accommodation may provide a powerful factor in determining oral proficiency as well as a criterion for interviewer training.
Discourse Validation of a Direct Versus a Semi-Direct Oral Test

Elana Shohamy

Establishing validity is an on-going process in which questions of different types are asked about a test. This paper illustrates the different phases needed in validation studies and the importance of studying tests from a variety of perspectives using the validation of two oral tests as an example.

Major attention has been given recently to the development of semi-direct oral tests (SOPI) as substitutes for more direct tests such as the Oral Proficiency Interview (OPI). In a number of studies the two tests correlated highly with one another (Stansfield and Kenyon 1989; Shohamy and Stansfield 1990). In a later study (Shohamy et al. 1991) the two tests were compared with regard to their discourse strategies and grammar showing that while the grammatical features of the two tests were the same, there were differences in a number of discourse strategies: paraphrasing-higher in SOPI; switch to L1-higher in OPI; error correction higher in SOPI.

This study is the third step in the validation study which examined whether the two tests, or the language samples obtained from them, were actually the same. Features known to exist in oral discourse were examined as to their emergence in the two tests. The specific discourse genre and a variety of discourse elements such as elaboration, expansion of ideas, switch of topics, and a variety of prosodic features, were examined and compared. Results showed that the two tests are not the same, they represent different discourse and contain different discourse elements. The language on the SOPI contains features of literate language and is similar in some ways to written language as the interlocutor is not present during the test. The OPI exemplifies conversational discourse features, typical of oral language that is contextualized, expanded, elaborated, etc. However, both tests do not include a large number of oral discourse features which are typical of oral language. The extent to which each of the tests is more construct valid (i.e., better represents the construct of oral proficiency), the implications of the results to oral testing in general and the steps which should be included in validation studies will be addressed.

Comparing the Scaling of Speaking Tasks by Language Teachers and by the ACTFL Guidelines

Charles Stansfield, Dorry Kenyon

Do classroom language teachers perceive the level of ability required to perform speaking tasks of various complexity in a manner compatible with the level of ability required for those tasks as outlined in the Guidelines for Speaking Proficiency developed by ACTFL? This paper compares the scaling of 38 speaking tasks in terms of the level of ability required to perform each one as perceived by language teachers in public schools with the scaling of the same speaking tasks according to the ACTFL Guidelines.

700 language teachers in Texas were randomly surveyed and requested to rate 38 foreign language speaking tasks, ranging from "Introduce Yourself" to "Discuss a Professional Topic", on a scale of 1 to 5 in answer to the question "Is the level of ability required to perform this task needed by language teachers in Texas public schools?" Responses were received from 62 French teachers, 121 Spanish teachers, and 240 bilingual education teachers. Data was scaled using the many-faceted Rasch analysis.

This paper presents the results of this survey by teacher group, and compares those results with the a-priori scaling of the tasks according to the ACTFL Guidelines. Similarities in scaling of the tasks support the conclusion that the ACTFL Guidelines represent a proficiency hierarchy perceived even by language teachers untrained in the proficiency movement.
The Impact of TWE Reader Qualifying Criteria on Reader Performance

Carol Taylor, Gerald DeMauro, Barbara Voltmer, Bruce Henderson

The test of Written English (TWE) program requires that all potential TWE readers demonstrate their ability to apply established scoring criteria accurately and consistently to TWE essays either by qualifying through a one-day reader training session or by reading successfully for other ETS essay reading programs.

The purpose of the present study is to document current TWE reader training procedures and evaluate reader performance data from TWE reader qualifying sessions and subsequent TWE readings to determine:

— how qualifying performance relates to later reader performance
— which variables examined from qualifying performance are the best predictors of successful reader performance
— whether or not decision points for qualifying and evaluating readers need to be adjusted.

The current measures of TWE reader performance include the following:

— number of essays read as first and second reader
— correlation of scores with those of other assigned readers
— discrepancy with other assigned readers
— ratio of standard deviations of scores
— mean score differences.

The measures of trainee qualifying performance include comparative statistics of the agreement between trainees and expert readers.

Multiple correlations and discriminant analyses, respectively, will be used to evaluate the relationships among the reader training and performance variables and to determine training variable values that best identify qualified readers. Samples included 619 reader trainees and 275 readers. Preliminary analyses suggest that current reader evaluation criteria enable the program to identify competent readers. Further descriptive analyses will be used to explain the outcomes of the study.
Validation of Measures of Grammatical Knowledge and Communicative Ability

Carolyn Turner, Jack Upshur

This paper will report the results of a construct validation study of 14 tests of grammatical knowledge (GK) and communicative ability (CA). This is the second phase of an investigation of two competing theories of second language learning. The two theories are: 1) attainment of grammatical knowledge in a second language precedes the development of communicative ability; 2) communication with limited linguistic resources provides the necessary condition for the development of grammatical knowledge. The study is being conducted in three phases. The final phase will be a cross-lagged time series study of GK and CA.

The first phase involved construction of tests of GK and CA utilizing eight different test methods. Fourteen of the 16 tests were tried out for feasibility, and appropriate modifications were made as a result of that trial. The rationale for the project and the results of the first phase were reported at LTRC 1991.

The second phase of the study involves the construct validation of seven tests each of two hypothesized second language traits, namely GK and CA. The 14 tests will be administered to a sample (n approximately 120) of learners from the same population which will be sampled for the final phase of the project. These are French speaking grade 5 students in an intensive ESL program in the greater Montreal area. Two different organizations of the test results will be analyzed. A multitrait-multimethod matrix of two traits and seven methods will be analyzed according to the original Campbell-Fiske criteria and by means of LISREL. Assumptions required for identification will be noted. The tests’ results will be organized also according to traits and facets of method as latent variables. These data will also be analyzed by means of LISREL.

We will report the MTMM matrix and summarize the findings about validity. We will also report the LISREL results and the structural models tested. Any differences between MTMM and facets-of-method results will be discussed. We will be able to report whether we have convergently and divergently valid tests of GK and CA which will allow continuation into the third phase of the study.
A Communicative Test in Analysis

Ingrid Wijgh

Final examinations for second language reading comprehension in the Netherlands consist of four long texts and about 40 multiple choice questions and of a so called "communicative part". In the communicative part there is a large variety of types of texts. In general the texts are short, authentic, printed in the original lay-out, and questions are based on the communicative approach.

A research project has been conducted to validate the communicative part, that is to answer the following questions:

1. In authentic texts learners will have to deal with a large number of unknown words. Does this affect their performance in a negative way?
2. Are the strategies used in answering the questions task-based or learner-based?
3. Are learners able to choose the most efficient strategies?
4. Do learners use information provided by the context?

Protocol-analysis (adapted version of Elshout's 7-steps method) was used to gather data on learners behaviour. Thirteen items of reading comprehension were given to 13 learners coming from two different schools. Data were collected on tape and analyzed afterwards. The main findings were: that learners use one basic strategy which is not always the most efficient one, that they do use the information provided by the context but only as a second choice and that unknown words do not bother them.

Protocol analysis also provided a lot of information about the test items. This method could be very useful as part of test development construction.

Discourse Variation in Oral Proficiency Interviews

Richard Young, Michael Milanovic

In this paper, a theoretical model of dyadic native speaker/non-native speaker (NS/NNS) discourse is described in terms of three features: interactional contingency, goal orientation of participants, and dominance. The model is then used to study the discourse of 30 dyadic oral interviews of the Cambridge First Certificate in English (FCE) examination.

The results of the study demonstrate the effectiveness of the model in abstracting the structure of oral interview discourse. They show that the discourse of oral proficiency interviews is characterized by greater reactivity by NNS candidates and greater orientation toward goals by NS examiners. Variation in the structure of the discourse is also investigated in this study. This is shown to be related to the examiner, the theme of the interview, the task in which the participants are engaged and the gender of the examiner and candidate.
Posters

TSE Validation in Federal Government Language Programs

Eduardo Cascallar, Marijke Walker

This is a validation study of the TSE in a government setting. This validation was carried out against ratings in the current version of the Federal Interagency Language Roundtable (FILR) Oral Proficiency Interview, and against supervisors’ evaluations of the participating government employees, as regards their speaking proficiency on the job. Reliability of the ratings between the government raters, and between these raters and ETS scores of the same tapes was also analyzed. The analyses also take into account several other background variables. In particular, it was important to study the validity of the TSE to evaluate the functional foreign language speaking ability in this new context, and to obtain an accurate and immediate indication of language proficiency, as well as to suggest what type of further training might be required.

A total of 120 non-native English speaking adult subjects was evaluated with the OPI and the TSE. Reliability of the ratings between the trained government raters, and between the government raters and ETS ratings of the same tapes were obtained. The TSE tapes were double-scored in each site, using a counterbalanced design so that each rater was paired with each other rater on the same number of tapes. Validity was examined using the OPI scores obtained from each examinee. As a second component of the validation study, supervisors of the examinees were asked to evaluate the participating government employees as regards their speaking proficiency on the job. This latter measure was also analyzed together with the TSE and the OPI scores in order to determine the relationship between all of these measures and additional background information including demographic, linguistic, and attitudinal data.

Evidence of a strong positive correlation (r = .79) between TSE (Overall Comprehensibility) scores and OPI ratings has been found. This was achieved without the personnel and logistical costs of the traditional OPI testing. In addition, diagnostic scores for pronunciation, grammar, and fluency were also included in the assessment. Overall, the TSE appears to be a good indicator of the adequacy of the examinees’ spoken English proficiency for those situations that have been examined. Implications for further test development, "special purpose oral proficiency testing", and training, will be discussed, as well as detailed results and conclusions.
Validity of Written Indicators of Second Language Proficiency

Alister Cumming, Dean Mellow

This cross-sectional study sought to identify linguistic features which validly indicate second language development, considering factors which often threaten the validity of measures of language proficiency, including the limitations of holistic or general measures, of accuracy-based measures, and of variations according to modality (written vs. oral), genre, and first language. The study analyzed 113 compositions of various genres (letter, argument, summary, cause/effect, description) written by adult students (24 Canadian Francophones and 45 Japanese visiting Canada) in intensive ESL programs. Specific measures such as TOEFL, oral interviews, class placements, and self-reports were used to categorize subjects within each first language group. Texts were analyzed for four features: lexical articles, plural -s nouns, 3rd person s on verbs, and a type/token ratio of different words. Preliminary multivariate analyses indicate that the construct of proficiency, as operationalized accounts for variance in accuracy of article use for both language groups (French, F = 9.6, p < .006; Japanese, F = 4.6, p < .03), but not for the other 3 dependent variables. Further sub-analyses of article and plural -s use are being conducted for hypothesized developmental stages (Pienemann and Johnston 1987; Master 1987; Peyton 1990). Genre differences account for some variation, interacting with article use, plural -s markers, and type-token ratios.

Hypothesis Testing in Construct Validation

Sara Cushing, Brian Lynch

Construct validation in language testing has frequently been investigated through correlational techniques such as factor analysis and multidimensional scaling of test data to discover underlying traits of test takers (e.g., Davidson 1988; Oltman and Sticker 1990). Another approach to construct validation is to use test data to test hypotheses about test taker attributes and their relationship to test scores. For example, one might hypothesize that students whose experience with English comes largely from a foreign language classroom might score better on a discrete-point test of grammar but worse on a test of extended listening than immigrant students who have learned English as a second language in an English-speaking country.

UCLA’s recently revised English as a Second Language Placement Examination (ESLPE) provides a rich source of data for such hypothesis-testing. The test consists of three subtests — Listening/Notetaking, Reading/Vocabulary, and Composition — which attempt to mirror the academic language skills required of university students. This paper describes efforts taken to validate the construct of “Academic Language Proficiency” (ALP) through hypothesis testing. The test, along with a discrete-point grammar test from an earlier version of the ESLPE, will be given to approximately 600 UCLA students in Fall 1991, who fall roughly into two categories: undergraduate students who were immigrants, and graduate students who are in the United States on student visas. Data from these tests will be used to investigate hypotheses such as the one stated above in an attempt to come to a better understanding of ALP and how it can be tapped in a university placement examination.
Canadian Test of English for Scholars and Trainees: CanTEST

Margaret Des Brisay, Jennifer St. John

The Canadian Test of English for Scholars and Trainees (CanTEST), in its English version and the French version, the Test pour étudiants et stagiaires au Canada (TESTCan), is a sophisticated test of language ability originally developed for Chinese candidates entering a Canadian academic and/or work environment. The CanTEST item bank has been used to compile tests for use in Canadian-funded projects in Indonesia and for use as an admissions test at several Canadian universities. A new version for short-term trainees has also been developed.

To date, the CanTEST has been successfully used with over 5,000 candidates. The CanTEST includes measures of four language skills (reading, writing, speaking and listening). Scores are reported in each of the four skills areas using a "band system" that relates test scores to a descriptive statement about the candidate's ability. Special features of the CanTEST include a test of skimming and scanning skills.

The CanTEST has been specifically designed to be responsive to the training objectives of Canadian-funded overseas language centres, and to the requirements of individual aid programs. As well, material is chosen for its compatibility with the educational background and world knowledge of those taking the test.

The poster session will focus primarily on two sub-tests which illustrate the special features described above. Presenters will also deal with questions concerning the suitability and availability of the CanTEST versions for other programs.

CHEMSPEAK: An Update

Dan Douglas, Larry Selinker

CHEMSPEAK, a specific purpose test of oral English ability for teaching assistants in chemistry, was reported on at the 1991 LTRC. It was suggested, based on somewhat limited data, that there may be a measurement advantage to using a field-specific language test, when the purpose is to make field-specific judgements. Since then, more data have been collected and a clearer picture of test performance has emerged.

In the poster session we will report on the latest analysis of results and elaborate on a procedure for constructing such field-specific tests, which primarily involves the manipulation of method facets, and a validation procedure which includes an interlanguage analysis of subjects' test performance.
Multiple-Choice Summary: A Measure of Test Comprehension

Ari Huhta, Elina Randell

The study aimed at finding an easily administrable alternative to summaries, open-ended questions, and other test types which purport to measure the comprehension of the main points in a text. Finnish LSP teachers, who would like to test the understanding of main ideas, need such a method because often they have to test several hundred students at a time and give out the results within only a few days. Valencia and Pearson (1987) suggest that the multiple-choice summary (a text accompanied by alternative summaries one of which is the best one) might be a good reading test. However, there appears to be no studies on the matter.

About 300 students (humanists and social scientists) at the University of Jyväskylä took a reading test which consisted of conventional multiple-choice and open-ended questions, a summarization task, and a multiple-choice (MC) summary. The students also filled in a questionnaire with questions about, for example, perceived sources of difficulty in reading and the face validity of the tests and texts.

The design of distractors in the MC summaries was based on features characteristic of poor (conventional) summaries.

The hypothesis was that those who do well on the tasks and questions aimed at understanding the main ideas should be the same ones who choose the right alternative in the MC summaries. Also, those who claim to be good at finding the main ideas in a text should be the same ones who make the right MC choice. An analysis of variance suggests that these hypotheses were partly correct.

The main problems with the new method is reliability. We tried to tackle the problem by asking the testees to justify their decisions in order to check whether they were just guessing. These justifications, however, seem difficult to interpret.

Implementation Issues: National Dutch Language Certification Examinations

Jan Hulstijn

In 1991, the Dutch Minister of Education appointed a committee to advise him on the establishment of a system which would make it possible for adult non-native speakers of Dutch to acquire language certificates, and provide information on the anticipated civil effect (the degree of acceptance) of such certificates in Dutch society. The committee report, published in November 1991, was well received. The poster will give details about the recommendations and the implementation process currently underway.
TOEFL 2000: Future Directions

Jacqueline Ross, Susan Chyn

Since 1976 the Test of English as a Foreign Language (TOEFL) has been a three-section, multiple-choice test consisting of test items designed to assess listening comprehension, structure and written expression, and vocabulary and reading comprehension.

While there is a notable body of research supporting the validity and reliability of the TOEFL exam as it exists today, there is also a concomitant recognition that the exam might be improved. Program staff and external committee members are exploring ways to refine the TOEFL within certain constraints. This poster session will present program needs and test development activities directed toward the design of the TOEFL as it might be presented in the future.

The presenters will actively invite audience participation and encourage attendees to offer suggestions as to how English proficiency might best be evaluated in the future to be of most benefit to all of TOEFL’s constituents, examinees, instructors and score users.

The A-LINC Assessment

Helen Tegenfeldt, Virginia Monk

Immigration Canada is starting long-range planning for a future nationwide program to provide language training for immigrants, which they have termed Language Instruction for Newcomers to Canada (LINC). Part of the planning process includes developing an assessment to determine who needs the language training, and how much of the sponsored language training is appropriate for that person. This assessment will be used in Employment and Immigration Centres (EIC) across Canada, and since it will be used to access the LINC program, it is being called the A-LINC.

The A-LINC targets only lower levels of English ability, from pre-literacy to pre-intermediate; it is intended to be given to non-ESL professionals; it takes from 10-20 minutes depending on the level of the client; it is more of a decision-making process than a formal test, and is comprised of a series of tasks that an EIC client is asked to do. At various points throughout the interview the counsellor may decide the client has reached the limit of this English ability and stops the interview, assigning the client to the appropriate level of language training associated with that exit point.

An earlier research project developed the basic form of this assessment, and tested the bottom and top exit points specifically. The current research project has added more exit points, based on EIC requirements, which are being tested on EIC clients, using both ESL and non-ESL interviewers; results will also be correlated with a language training program’s assessment of these clients, just prior to their entering class, and then following up with teachers’ assessments after a month of class.
The purpose of this poster session is to make the OTESL materials available for inspection by interested persons in the field. This battery of performance-based EAP instruments includes a 45-minute Placement Test with listening, reading, writing and speaking components, specifically designed for use in intensive pre-admission ESL programs, as well as a 2 ½ hour Post-Admission Test which measures performance in reading, listening and writing, and an Oral Interaction Test. The latter instruments are designed for students already admitted to academic programs who may still need further English instruction. They measure global proficiency levels according to assessment scales based on functional descriptions, and provide diagnostic feedback on speaking and writing. Alternative science and non-science versions of these two tests are available. The Oral Interaction Test is also noteworthy for the task difficulty hierarchy and adaptive format.

The OTESL instruments were developed from detailed specifications using authentic source materials to reflect academic discourse types, topics and tasks which simulate the language use needs identified for ESL speakers in North American academic situations. The development process and the instruments themselves reflect the considerable advantages as well as the logistical drawbacks of a theme-based performance approach to EAP testing. The test manual provides information on validity and reliability. The OTESL, developed by a team of researchers for the Ontario Ministry of Colleges and Universities from 1983-88, was published by the Joint Language Testing Service of the Universities of Ottawa and Toronto in 1990.
Applications of TESTGRAF in Setting Cut-off Points on ESL Tests

Margaret Des Brisay

The presenter will show both colour monitor displays and hard copies of the output from the TESTGRAF analysis of a multiple choice ESL test battery.

Developed by Dr. Jim Ramsay of McGill University, TESTGRAF provides all the usual indicators of multiple choice item performance including 3 parameter IRT estimates. There are two special features of this program which make it an attractive addition to the present set of software options for language testers. First of all, all output from the analysis including option characteristic curves, test information functions, and a principal components analysis of correct option item characteristics, can be graphically displayed on a colour monitor (or printed out on a postscript printer), making the information easier to assimilate. Secondly, and more importantly, TESTGRAF makes use of the information to be obtained from the various wrong options which were chosen for incorrectly answered items. The output is displayed in "examinee credibility plots" so that, for each examinee, the likelihood or relative credibility of the true proficiency being at value 0 is shown.

Deciding whether an examinee scores beyond a certain cut point makes heavy demands on precision estimates. Using TESTGRAF, it may be possible to ensure that only people who are statistically rather than numerically below a cut point are rejected.

Using the Information Curve to Assess Language CAT Efficiency

Michel Laurier

Since a placement test is designed to assess students general proficiency on a broad ability range, adaptive testing is particularly appropriate for placement purposes. In our study, we developed a computerized adaptive placement test in French at the post-secondary level. This test seeks to measure abilities underlying various uses of the second language and is composed of three sub-tests. The whole test is administered using a micro-computer and is available with two different item selection algorithms both based on IRT procedures. The item banks were also used to create two conventional "paper and pencil" parallel versions.

As reliability indices are meaningless in an IRT context, we used the information curve to assess the CAT efficiency versus the conventional versions. Efficiency is defined as the ratio between the information that is obtained and the number of items that have been presented. The study shows that the adaptive form uses much fewer items than the "paper and pencil" form to achieve the same reliability. However data obtained from post-secondary students using one form as a pre-test and the other one as a post-test show that the test scores cannot be compared because the marking procedures are different.

CAT efficiency is often obtained at the expense of validity because of the constraints of the IRT framework and the technology limitations. Therefore one must be very careful using an adaptive procedure for language testing.
Rapid Profile: On-line Diagnostic Screening of Interlanguage Acquisition by Expert System

Manfred Pienemann, Ian Thornton, Alison Mackey

Rapid Profile is based on psychologically plausible and objectively measurable reference criteria obtained from second language acquisition research. In our paper we will show that the close relationship between SLA research and our screening procedures ensures a high degree of construct validity. We will also demonstrate that the procedure has a high degree of reliability.

Rapid Profile is a second languages screening procedure which allows the analyst to assess a learner's level of development against standard patterns in the acquisition of the target language. The procedure is based on a sample of natural speech which is rapidly analyzed with the help of a computer program. The procedure has evolved from detailed empirical studies of English as a second language which established fixed and psychologically plausible reference criteria for the screening process.

We will demonstrate the construct validity of Rapid Profile and deal with two major reliability aspects of the procedures: 1) the efficient elicitation of relevant data and 2) the level of accuracy obtained in the observation task through interactive training.

Firstly, we will demonstrate that developmental interlanguage features can be selectively triggered by specific communicative tasks. In an experiment we analyzed the spontaneous speech produced by ESL learners and native speakers in response to a set of communicative tasks. The morpho-syntactic structures produced by the subjects varied predictably according to task. This predictability can be used to increase the speed and reliability of our procedure.

Secondly, we will demonstrate that the accuracy of Rapid Profile analysis is comparable to full linguistic profiling. We trained 10 analysts using an interactive audio CD technique which realistically simulates the observation task. The training software recorded and analyzed the performance of the analysts.
Symposia

The Educational and Social Impact of Language Tests

As validity is the main theme of the 1992 LTRC Colloquium, the symposium will address the expanded view of validity, that which is adopted by Messick (1988, 1989) and others whereby construct validation includes also evidence of tests' use. Specifically, the utilization and consequences of test results by decision makers, the impact that tests have on test takers, teachers, and on the educational and social system in which they operate. Accordingly, the role of the tester does not end in the development phase but needs to examine issues of test use, relevance, ethics and the interpretation of scores by decision makers and students. Thus, "testers cannot be viewed any more as technicians who complete their work when they reach satisfactory reliability coefficients, instead they must also consider the consequences, social, psychological, ethical, curricular, and educational of the tests which they produce."

Introduction

Elana Shohamy, Eduardo Cascellar

The introduction provides the framework through which to view the three papers and outlines the ethical issues raised by language testing.

Does Washback Exist?

Charles Alderson and Diane Wall

The concept of backwash, or washback as it has been named in the language testing literature, i.e., the influence of a test on teaching, is commonly encountered in textbooks on language testing, and its existence, importance and influence is frequently asserted, and very occasionally exemplified by anecdotes. However, a review of the use of the term washback reveals that the concept is rarely defined in the literature, and even less commonly investigated empirically. If washback is a powerful phenomenon as is asserted, then it would seem important to begin to understand what exactly it consists of, what effects if brings about and how. If we are to attempt to engineer beneficial washback with our tests, or even if we wish to use our tests as "levers for change" (Pearson 1988), then we need to define what sort of influence tests in general and tests specifically might and do have on teaching and the curriculum.

This paper is intended to initiate a debate about what washback might consist of, what research has been conducted to date into the "phenomenon" and how as language testers we might devise a series of studies to examine the alleged washback of our tests.

The Power of a Test: A Study on the Effect of a Reading Comprehension Test on Language Learning

Elana Shohamy

The issue of test use has gained major attention recently with the increased use and introduction of new tests as devices aimed at solving major educational problems. There is a belief by educators, politicians and administrators that the power and authority of tests, per se, will be instrumental in saving and upgrading declining academic achievements. Although a number of studies have been conducted recently to investigate the impact that tests have on learning, instruction and the school social structure (Smith 1991; Shepard 1991), not much is known about the impact and use of language tests.
In this paper results of a study that examines the impact of newly introduced language tests on language learning is reported. The main focus will be on a reading comprehension test administered nationally to 5th and 6th grade students, but reference will be made to other language tests. Data was collected through observations of classes, teaching material and other documents produced before and after the test administration (e.g., report cards, lesson plans, teacher notebooks, Ministry of Education notes and reports, and questionnaires and interviews with students, teachers and parents).

Results indicate that tests have a strong impact on learning. This impact is mostly instrumental and not conceptual, and varies according to a number of variables such as the purpose of the test, whether the group failed or passed, the status of the principals, etc. The nature of the impact will lead to conclusions regarding possible solutions for integrating tests with instruction, for increased responsibility of testers as to the use of the tests, for sharing testing discourse with the users, and for less use of tests for power and control.

Exercising Washback: The Sri Lankan Impact Study

Diane Wall, Charles Alderson

This paper is intended to contribute to the investigation of washback and its alleged effects by reporting on the results of a four-year project looking at a particular public examination in a Third World context, and examining the impact of innovations in language tests on teachers' and pupils' behaviours in the classroom. The paper reports on the use of a variety of classroom observation instruments and questionnaires and on the content analysis of a number of teacher-made tests in an attempt to establish the degree to which new tests can be said to have influenced teaching.
The Development and Use of Rating Scales in Language Testing

The four papers in this symposium discuss research issues in the development and application of scales to rate linguistic performance. The papers highlight theoretical issues and illustrate empirical techniques for research use. The objective of the symposium is not only to transmit information, but also to provide a forum for the discussion of issues involved in using raters to judge linguistic performance. As measurement of language becomes more holistic and the use of performance assessments in the measurement field as a whole increases, we believe that scale development will become increasingly important in language testing research.

Introduction

Dorry Kenyon

The introduction sets the stage for the presentation of the papers. Issues common to all the papers are highlighted and a definition of terms is provided.

Direct Scalar Rating vs. Multiple Binary Rating of ESL Learner Writing

Frank Davidson, Liz Hamp-Lyons, Dorry Kenyon

How can raters be encouraged to consider (and use) an entire rating scale rather than "favourite" scores? Multiple Binary Rating (MultiBin) has been suggested as an alternative to Direct Scalar Rating (DSR) as a means to avoid discrepancies in scale-step probabilities. This paper reports on the analysis of data sets from two operational settings in which binary procedures have been attempted. The analysis of qualitative data collected from raters using both DSR and MultiBin is also presented. Discussion focuses on the advantages and disadvantages of MultiBin over DSR.

Validation of a New Holistic Rating Scale Using Rasch Multi-Faceted Analysis

Belle Tyndall, Dorry Kenyon

What happens to the accuracy of placing students within a college EFL program when an analytical scale is replaced by a holistic essay rating scale in which the scale points reflect the program levels? This paper discusses the development and analysis of such a scale. A multi-faceted Rasch analysis is used to investigate how the scale was applied by 11 raters scoring over 200 papers. Correlational analysis and an analysis of misplacements provides further evidence of the new scale's validity. Feedback from raters on how they applied the scale is also discussed.

Developing Rating Scales for the CASE: Theoretical Concerns and Analyses

Michael Milanovic, Nick Saville, Anne Cook

What are the relationships between the development of a rating scale and models of communicative competence, practical considerations of testing procedures, test tasks, and the pool of test raters and test takers? What are the appropriate methods to use to determine appropriate scale points? This paper uses the development of the Cambridge Assessment of Spoken English (CASE, presented at a poster session at LTRC 1991) to illustrate 1) the need to treat the development/validation of rating scales as an integral part of the overall test development process and 2) the analysis of rating scales. Analyses of ratings obtained from "live" pilot administrations of the CASE using Rasch techniques are reported.
Rating Scales and Native Speaker Performance on a Communicatively Oriented EAP Test

Tim McNamara, Jan Hamilton, Eileen Sheridan

What are the implications for the use of native speaker performance as a reference point (explicit or implicit) in scalar descriptions of second language speaker performance? What are we actually testing in communicatively oriented tests of English for academic purposes? This paper examines the construct validity of such tests in the light of a discussion of the concept of "performance testing" in ESP contexts. A distinction is proposed between a strong and a weak sense of the terms "performance test". The extent to which EAP tests are simultaneously tests of language proficiency and other skills (e.g., study skills) is considered. The results of two studies of the performance of native speaker on the IELTS test are examined in the light of the above distinction.

Summary

Dorry Kenyon
Addresses of Program Participants

Charles Alderson  
Dept. of Linguistics & Modern English Language  
Bowland College  
University of Lancaster  
Lancaster LA1 4YT  
United Kingdom

Annette Cook  
University of Cambridge  
Syndicate Buildings  
1 Hills Road  
Cambridge CB1 2EU  
ENGLAND

Kathleen Bailey  
Monterey Institute of International Studies  
426 Van Buren Street  
Monterey, CA 93940

Alister Cumming  
Ontario Institute for Studies in Education  
Modern Language Centre  
252 Bloor Street W.  
Toronto, Ontario M5S 1V6

Anne Brown  
Language Testing Centre  
Dept. of Linguistics and Language Studies  
University of Melbourne  
Parkville Victoria 3052  
Australia

Sara Cushing  
TESL Centre & Applied Linguistics  
U.C.L.A.  
3300 Rolfe Hall  
Los Angeles, CA 90024

James Dean Brown  
Department of ESL  
University of Hawaii  
1890 East-West Road  
Honolulu, HI 96822

Fred Davidson  
DEIL, 3070 FLB, UIUC  
707 S. Matthews  
Urbana, IL 61801

Gary Buck  
Monterey Institute for International Studies  
425 Van Buren Street  
Monterey, CA 93940

Alan Davies  
University of Edinburgh  
Department of Applied Linguistics  
14 Buccleuch Place  
Edinburgh EH8 9LN  
Scotland, U.K.

Eduardo Cascallar  
Educational Testing Service  
Rosedale Road, Mail Stop 10-P  
Princeton, NJ 08541-0001

John de Jong  
CITO  
P.O. Box 1034  
6801 Mg Arnhem  
The Netherlands

Susan Chyn  
Educational Testing Service  
P.O. Box 6155  
Princeton, NJ 08541

Gerald DeMauro  
Educational Testing Service  
Mail Stop 30-P  
Rosedale Road  
Princeton NJ 08541

Caroline Clapham  
Dept. of Linguistics  
Lancaster University  
Bailrigg  
Lancaster  
LA1 4YT England

Margaret Des Brisay  
Second Language Institute  
University of Ottawa  
600 King Edward  
Ottawa, Ontario  
K1N 6N5
Emily Detmer  
Dept. of English as a Second Language  
University of Hawaii  
1890 East-West Road  
Honolulu, Hawaii 96822

Thom Hudson  
Department of ESL  
University of Hawaii  
1890 East-West Road  
Honolulu, Hawaii 96822

Dan Douglas  
1105 Brookridge  
Ames, IA 50010

Ari Huhta  
University of Jyva Skyla  
Language Centre for Finnish Universities  
P.O. Box 35  
SF-40351 Jyva Skyla  
Finland

Catherine Elder  
Language Testing Centre  
Dept. of Linguistics and Language Studies  
University of Melbourne  
Parkville Victoria 3052  
Australia

Jan H. Hulstijn  
Applied Linguistics Department  
University of Amsterdam  
De Boelelaan 1105  
1081 HV Amsterdam  
Netherlands

Janna Fox  
Centre for Applied Language Studies  
215 Paterson Hall  
Carleton University  
Ottawa, Ontario  
K1S 5B6

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

Claire Gordon  
Department of English  
The Open University of Isreal  
16 Klausner Street  
Ramat Aviv, Tel-Aviv 61392  
Israel

Michel Laurier  
Université de Montréal  
Faculté des sciences de l’éducation  
C.P. 6128, succursale A  
Montréal (Québec)  
H3C 3J7

Jan Hamilton  
Language Testing Centre  
Dept. of Linguistics and Language Studies  
University of Melbourne  
Parkville, Victoria 3052  
Australia

Tom Lumley  
Language Testing Centre  
Dept. of Linguistics and Language Studies  
University of Melbourne  
Parkville, Victoria 3052  
AUSTRALIA

Liz Hamp-Lyons  
Department of English and Applied Linguistics  
Campus Box 175, P.O. Box 173364  
University of Colorado at Denver  
Denver, CO 80217-3364

Denise Lussier  
Faculty of Education  
McGill University  
3700 McTavish  
Montréal, Quebec H3A 1Y2

Grant Henning  
Educational Testing Service  
Mail Stop 10-P  
Rosedale Road  
Princeton NJ 08541

Brian Lynch  
P.O. Box 1926  
Beverly Hills, CA  
90213 U.S.A.
Alison Mackey
L.A.R.C.
University of Sydney
Sydney, Australia
NSW 2006

Tim McNamara
Language Testing Centre
Dept. of Linguistics and Language Studies
University of Melbourne
Parkville, Victoria 3052
Australia

Dean Mellow
University of British Columbia
Centre for the Study of Curriculum and Instruction
2125 Main Mall
Vancouver, B.C.  V6T 1Z4

Michael Milanovic
University of Cambridge
 Syndicate Buildings
1 Hills Road
Cambridge CB1 2EU
ENGLAND

Virginia Monk
Vancouver Community College
Box N 24620 Station "C"
Vancouver, B.C.
V5T HN4

Kyle Perkins
Linguistics Department
Southern Illinois University
Carbondale IL 62901

Manfred Pienemann
L.A.R.C.
University of Sydney
Sydney, Australia
NSW 2006

Don Porter
Welwyn Garden City
Herts, England
AL8 7PW

Timothy Pychyl
Department of Psychology
Carleton University
Ottawa, Ontario K1S 5B6

Elina Randell
University Jyva Skyla
Language Centre for Finnish Universities
P.O. Box 35
40351 Jyva Skyla
Finland

Doreen Ready
University of Ottawa
Second Language Institute
600 King Edward Avenue
Ottawa, Ontario
K1N 6N5

Jacqueline Ross
Educational Testing Service
P.O. Box 6155
Princeton, NJ 08541

Steven Ross
Department of English as a Second Language
University of Hawaii at Manoa
1890 East-West Road
Honolulu, HI 96822

Nick Saville
University of Cambridge
Local Examinations Syndicate
Syndicate Buildings, 1 Hills Road
Cambridge CB1 2EU
England

Larry Selinker
ELI, University of Michigan
Ann Arbor, MI
48109 U.S.A.

Eileen Sheridan
Language Testing Centre
Dept. of Linguistics and Language Studies
University of Melbourne
Parkville, Victoria 3052
Australia

Bruce Shields Henderson
Santa Clara University
Room 223, St. Joseph's Hall
Santa Clara, CA 95033
Elana Shohamy  
Tel Aviv University  
School of Education  
Tel Aviv, Israel  
69978

Belle Tyndall  
Department of English as a Foreign Language  
Academic Center, T607  
The George Washington University  
Washington, D.C.  20052

Peter Skehan  
53 The Avenue  
Muswell Hill  
London WC1H  
ENGLAND

John Upshur  
TESOL Centre  
Condordia University  
1455 Le Maisonneuve Blvd. West  
Montreal Quebec H3G 1M8

Bernard Spolsky  
Department of English  
Bar Ilan University  
52-100 Ramat-Gan  
Israel

Barbara Voltmer  
Bay Area Essay Reading Office  
Educational Testing Service  
Emeryville, CA

Jennifer St.John  
Second Language Institute  
University of Ottawa  
600 King Edward  
Ottawa, Ontario  
K1N 6N5

Marijke Walker  
Federal Bureau of Investigation  
Language Services Unit, Room GRB 210  
Washington, D.C.  20535

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

Dianne Wall  
I.E.L.E.  
University of Lancaster  
Lancaster LA1 4YT  
ENGLAND

Carol Taylor  
International Testing and Training Programs  
Educational Testing Service  
Princeton, N.J.  08541

Marjorie Wesche  
Second Language Institute  
University of Ottawa  
600 King Edward Avenue  
Ottawa, Ontario  
K1N 6N5

Helen K. Tegenfeldt  
Vancouver Community College  
1155 E. Broadway  
Box 24620 Station C  
Vancouver B.C. V5T 4N3

Ingrid F. Wijgh  
Numankade 8  
Utrecht, Holland  
3572 KZ

Ian Thornton  
L.A.R.C.  
University of Sydney  
Sydney, Australia  
NSW 2006

Richard Young  
Department of Linguistics  
University of Southern Illinois at Carbondale  
Carbondale, Illinois  62901-4517

Helen K. Tegenfeldt  
Vancouver Community College  
1155 E. Broadway  
Box 24620 Station C  
Vancouver B.C. V5T 4N3

Carolyn Turner  
Faculty of Education  
McGill University  
3700 McTavish  
Montreal, Quebec  
H3A 1YA

Bruno D. Zumbo  
Faculty of Education  
University of Ottawa  
Ottawa, Ontario  K1N 6N5