TENTH ANNUAL

LANGUAGE TESTING

RESEARCH COLLOQUIUM

University of Illinois at Urbana-Champaign

March 5-7, 1988
TENTH ANNUAL
LANGUAGE TESTING
RESEARCH COLLOQUIUM

SPONSORED BY

The Division of English as an International Language
and
The Office of International Programs and Studies
University of Illinois at Urbana-Champaign

Conference Organizers

Lyle F. Bachman
University of Illinois at Urbana-Champaign

Adrian S. Palmer
University of Utah

With Assistance From

Cindy Meyer-Giertz

Rosa Shim
ACKNOWLEDGEMENTS

Institutional sponsorship for this conference has been graciously provided by the Division of English as an International Language and the Office of International Programs and Studies, of the University of Illinois at Urbana-Champaign. Additional support has also been generously provided by the following campus units:

Department of French
Department of Germanic Languages and Literatures
Office of Instructional and Management Services
Language Learning Laboratory
Department of Slavic Languages and Literatures
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chronology of the Language Testing Research Colloquium</td>
<td>3</td>
</tr>
<tr>
<td>Schedule of Activities</td>
<td>4</td>
</tr>
<tr>
<td>Abstracts of Conference Presentations</td>
<td>7</td>
</tr>
<tr>
<td>(Appearing in Alphabetical Order by Speaker)</td>
<td></td>
</tr>
<tr>
<td>Symposium on Self-Assessment</td>
<td>24</td>
</tr>
</tbody>
</table>
PROGRAM INTRODUCTION

Nearly ten years ago a relatively small group of applied linguists from several countries, with varied backgrounds and perspectives, braved the rigors of the New England winter and the vagaries of convention planning, which forced them to squeeze into a single hotel room for two days, in order to participate in a discussion of the issues and problems of language testing. For many who were at that first colloquium, it was both surprising and gratifying to find that there were others who shared their interest in this abstract and relatively unpopular enterprise. Equally amazing was the discovery that not only did we share common professional interests, but we were in many other ways "kindred spirits". As Palmer, Groot and Troper (1981) observed in their forward to the volume of papers that came out of that first colloquium, "the colloquium has enabled people with a common narrowly-defined interest to get to know each other and to develop the closeness and the lines of communication that allow each to profit more fully from the work of others" (p. vii). That observation is as true today as it was then. Indeed, it is this spirit of camaraderie, of personal closeness, based on mutual respect, that has nourished the spirited annual debates of professional issues and year-around communication among colleagues who are also friends that have become the hallmarks of the Language Testing Research Colloquium.

The past ten years have seen many developments in language testing, and the Colloquium, with its focus on research, has provided the sounding board for many of these. The use of confirmatory factor analysis as an approach to the construct validation of language tests was forged by the debate and collective effort among the participants at the first Colloquium. Research in oral testing has been a frequent topic at the Colloquium, and it is safe to say that many refinements in the measurement of oral ability have been influenced by discussions at colloquia. More recently, the applications of item-response theory and multidimensional scaling to language testing research have provided fodder for the Colloquium's cannon, and have emerged all the stronger and more promising in the process. Not all of these developments have been chronicled, but the four published volumes that have come out of the Colloquium (Palmer, Groot & Troper 1981, Jones, DesBrisay & Paribakht 1985, Stansfield 1986, and Bailey, Dale & Clifford 1987) provide an overview of the issues that have captivated the field over the past ten years.

The first Colloquium came at a time when John Oller's research into the nature of language ability was redefining our view of language testing and at the same time raising questions that would lead to the emergence of language testing as a subfield of applied linguistics in its own right, with its own research questions, and with a research methodology that would contribute to other areas of applied linguistics. It also came at a time when Mike Canale and Merrill Swain were formulating the ideas that would emerge in their seminal paper on teaching and testing communicative competence. As a result of these cross-currents, it is not surprising that the focal points that emerged from the first Colloquium were an interest in a broader view of language proficiency as communicative competence, and a determination to embark on a program of empirical research into the then relatively unknown realm of construct validation.

The Colloquium has had a variety of themes over the years, and in some years has had no particular theme, other than a focus on research. We felt it was timely, in this tenth year of our persistence, to return to the theme of the first colloquium: the validation of tests of communicative language ability. Timely, because although "communicative" as a buzz-word has lost a certain amount of cachet in language teaching, it appears that language testers are beginning to come to grips with what the characteristics of "communicative" tests are. Thus,
language testing may offer one avenue for investigating the nature of both communicative language use and the very abilities that make such use possible. One question we may ask ourselves, then, is, "How far have we really come in the past ten years toward understanding the language abilities that we profess to measure?"

We might also look at technological developments as an indication of our emergence as a field. In 1979, technical sophistication focused primarily on research design and the analysis of results—Clifford's examination of multitrait-multimethod correlations and Engelskirchen, Cottrell and Oller's principal components analysis were "state-of-the-art". In the past four or five years, we have seen the increasing application of technology to test design and administration, and we now have at our disposal not only a wider range of analytic tools, but more powerful ones as well. Appropriate questions to ponder in this regard are, "How much have these technical advances contributed to our understanding of the fundamental issues of language testing?" and "Are we simply probing the same questions in greater detail, or are we asking new questions?"

While the tenth annual Colloquium is perhaps no more special than was the ninth or than will be the eleventh, we feel this is an occasion to celebrate the remarkable "staying power" of the Colloquium. As Stansfield (1988) noted in his introduction to the papers from the seventh Colloquium, we have no charter, no officers, no dues. Every year there are new faces and new perspectives which add to the fabric of our collective identity. The Colloquium endures in spite of our resistance to becoming formally organized. It thrives because of our common interest in and commitment to the field of language testing, and because we truly enjoy our work, especially bashing heads once a year. And we have fun together. After all, some of our best friends are language testers.

Lyle F. Bachman

Adrian S. Palmer
CHRONOLOGY OF THE TESTING RESEARCH COLLOQUIUM

AND COLLECTIONS OF COLLOQUIUM PAPERS

First LTRC: Boston, TESOL, 1979


Second LTRC: San Francisco, TESOL, 1980

Third LTRC: Honolulu, TESOL, 1981

Fourth LTRC: Ann Arbor, University of Michigan, 1982

Fifth LTRC: Ottawa, University of Ottawa, 1983


Sixth LTRC: Houston, TESOL, 1984


Eighth LTRC: Monterey, Defense Language Institute, 1986


Tenth LTRC: Urbana, University of Illinois, 1988
SCHEDULE OF ACTIVITIES

Saturday, March 5, 1988

10:00 - 12:00 Registration
Illini Union Alumni Lounge

1:00 - 1:30 Opening Ceremonies
Room 209, Illini Union
Welcoming Remarks: Braj B. Kachru, Director,
Division of English as an International Language

2:00 - 2:45 Michael Canale: "The content validity of some oral interview
procedures: An analysis of communication problems and
strategies"

2:45 - 3:15 Health Break

3:15 - 4:00 Fred Davidson: "An exploratory modeling survey of the trait
structures of existing communicative language test datasets"

4:00 - 5:30 General Issues in Large-Scale Validation Studies

Charles Alderson: "New procedures for validating proficiency
tests of ESP? Theory and practice"

Charles Alderson, Fred Davidson and Dianne Wall: "Validating
tests in difficult circumstances"

Charles Stansfield and Jacqueline Ross: "A long term research
agenda for the Test of Written English"

Sunday, March 6, 1988

Room 275-279, Illini Union

9:00 - 9:45 Esin Kaya-Carton: "Empirical comparison of three item-
calibration methods in validating French reading proficiency
levels"

9:45 - 10:30 Gordon Hale, Charles Stansfield, Donald Rock, Marilyn Hicks,
Frances Butler Hinofotis and John Oller: "Multiple-Choice cloze
items and the Test of English as a Foreign Language"
10:30 - 11:00  Health Break

11:00 - 12:15  Characteristics of Test Takers and Test Performance

Peter Hargreaves and John Foulkes: "The relationship between candidates' subject specialization and their performance in a test of English for specific purposes"

Brian Lynch, Fred Davidson and Grant Henning: "Person dimensionality in language test validation"

Philip Oltman and Lawrence Stricker: "How native language and level of English proficiency affect the structure of the Test of English as a Foreign Language"

12:15 - 2:00  Lunch

2:00 - 2:45  Liz Hamp-Lyons, Grant Henning and Gerald DeMauro: "Construct validation of communicative writing profiles"

2:45 - 3:30  Dan Douglas and Spencer Swinton: "A study of validity characteristics of the SPEAK test"

3:30 - 4:00  Health Break

4:00 - 4:45  Lyle Bachman, Antony Kunan, Swathi Vanniarajan and Brian Lynch: "Task and ability analysis as a basis for examining content and construct comparability of two EFL proficiency test batteries"

6:00 - 7:00  Reception
General Lounge, Illini Union

Hosted by
The University of Cambridge Local Examinations Syndicate

7:30 -  
Banquet
University Inn, Century Room

Monday, March 7, 1988

Room 275-279, Illini Union

9:00 - 9:45  Aaron Carton: "Linguistic analysis of test passages as a method of determining construct validity"
9:45 - 10:30  John Clark: "Validation of a tape-mediated, ACTFL/ILR-scale based test of Chinese speaking proficiency"

10:30 - 11:00  Health Break

11:00 - 12:15  Business Meeting

12:15 - 2:00  Lunch

2:00 - 2:30  Elana Shohamy and Ofra Inbar: "Content and construct validation of listening comprehension tests"

2:30 - 5:00  Symposium on Self-Assessment Techniques in Language Learning: Implementation, Validity and Rationale

Organizer: Russanne Hozayin

I. An Overview of Self-Assessment in Language Learning

Mats Oskarsson: "Self-assessment of language proficiency: rationale and applications"

II. Research and Validity Issues

Bernard Spolsky: "Facilitating accurate self-assessment of functional language skills"

Anne-Mieke Janssen-van Dieten: "The relationship of self-assessment to external assessment measures of Dutch as a second language"

Russanne Hozayin: "The validity of self-assessment techniques in language learning: comparing ability and self-assessment of adult Egyptian EFL students"

Lyle Bachman and Adrian Palmer: "The construct validation of self-ratings of communicative language ability"

III. Application of Self-assessment in On-going Programs:

Leslie Dickinson and Gillies Houghton: "Collaborative assessment by master's candidates in a tutor-based system"

6:52  AMTRAK leaves for Chicago
NEW PROCEDURES FOR VALIDATING PROFICIENCY TESTS OF ESP?
THEORY AND PRACTICE

J. Charles Alderson
University of Lancaster

Within the world of ESP testing, be it proficiency or achievement, it has become relatively common practice to attempt to conduct some form of needs analysis, where the test population's likely future use of the language in question is identified. Specifications for test content and format are drawn up on the basis of such needs analyses. The derivation of particular test items and test forms is supposedly a simple matter of operationalising the specifications.

However, experience in attempting to apply such procedures has identified a number of more or less serious problems and issues:

1. The conduct and the content of the needs analysis itself may be judgemental or empirical. Armchair speculations about students' language needs may be easier than observation or data-gathering via questionnaires or interview procedures, but are at least a priori likely to be limited in validity. The design of suitable data-gathering instruments is itself, however, problematic if it is not to precondition the responses from informants. Open-ended procedures face the difficulty of obtaining uninformed or superficial accounts of target needs or of gathering widely-divergent responses which are then difficult to categorise.

2. Whatever the means of initially drawing up a description of target needs, test specifications of a limited finite nature then have to be established from an inevitably small subset of the needs identified. Problems of inadequate or biased sampling abound, and criteria for inclusion and exclusion of potential items, skills or abilities are often of dubious validity.

3. Once specifications of test content have been drawn up there is still the problem of turning these into test items, and it is at this stage that many practical difficulties are faced. Some specifications may not be operationalisable, whereas others may only be turned into trivial test items. More importantly, recent research has shown that there is often a serious gap between test constructors' intentions with respect to test content (i.e. the test specifications) and expert judgements of what particular test items are actually thought to measure. This may be due to poor specifications, poor item construction, or, more seriously, it may reflect an underlying difficulty in the relationship between test specifications and test validity.

In the development of a major international proficiency test, new approaches to test specification and construction are being tried, in order to avoid at least some of the above problems. In this case, experienced applied linguists are operationalising their insights into necessary test content in the form of draft specifications and related draft items. This phase is being carried out without an empirical needs analysis; however, in the second phase the specifications and items are then being taken to informants from relevant subject disciplines as well as to test takers, and opinions are being solicited as to the appropriacy of items and specifications and the relationship between the two. In addition, testee insights are being gathered into the nature of the processes they undergo when completing the test items. These latter insights will be related to the draft specifications upon which the items were supposedly based. As a result of this extensive inspection period, both items and specifications will be revised in order to result in guidelines for test constructors, the results of which will then be subject to more normal validation procedures. It is hoped that by this means test content validity might be more appropriately established. It remains to be seen, however, whether such innovative procedures will result in improved construct, concurrent or predictive validity.
VALIDATING TESTS IN DIFFICULT CIRCUMSTANCES

J. Charles Alderson
Dianne Wall
University of Lancaster

Fred Davidson
University of California at Los Angeles

The validation of tests of communicative language ability is both desirable and difficult, if not impossible. This is particularly problematic for the concurrent validation of tests claiming to be innovative, since any direct comparison with other measures of dubious construct validity must be uninterpretable. Other measures or estimates of validity - judgemental in particular - are possible but of arguably less value. Since the empirical aspects of construct validity are at least as problematic, for communicative tests as they are for noncommunicative tests, construct validation is also difficult.

Apart from theoretical difficulties, however, there are many testing situations which are under-resourced and where conditions do not favour experimentation with tests - where test validation in general, let alone validation of tests of communicative language ability, is a major practical problem. In such situations, opportunities for empirical pre-testing may be limited or non-existent, tests may have to be constructed anew for each administration, and the lack of security may mean that items cannot be calibrated across versions. Moreover, it may be impossible to administer more than one test to the population in question and the gathering of other data on candidates' abilities can be a logistical nightmare.

This paper will present such a situation - in Sri Lanka - and describe what attempts were made at the validation of an innovative test which, it was hoped and claimed, tested communicative language ability. After describing the context and purpose of the test in question, the paper will discuss the procedures for and the results of establishing content validity by judgemental means. A comparison between responses to questionnaires administered concurrently with the test and test results revealed limited but useful information on possible test bias and the test's relationship with other measures (no other empirical or judgemental measures of individual ability were possible). Internal analyses of test performance, including item analytic, correlational, factor analytic and multi-dimensional scaling approaches, will also be reported, although it will be suggested that their value is severely limited without independent data on the characteristics of the test population. The performance of two similar tests (supposedly parallel versions) over two years (since any one test version could only be administered once) will be compared to the first test for stability of test construct, if not its validity.

The paper will conclude with a discussion of the practical and theoretical difficulty of such attempts at validation and call for further considerations of and research into possible solutions to the validation issue as a matter of urgency.
ABILITY AND TASK ANALYSIS AS A BASIS FOR EXAMINING
CONTENT AND CONSTRUCT COMPARABILITY
IN TWO EFL PROFICIENCY TEST BATTERIES

Lyle F. Bachman
University of Illinois at Urbana-Champaign

Antony Kunan
Swathi Vanniarajan
Brian Lynch
University of California at Los Angeles

A common problem in language testing is that of attempting to determine the "comparability" of different tests. Although test users sometimes identify comparability with the simple equivalence of scores, the examination of test comparability must begin with an assessment of the extent to which tests measure the same abilities, and hence must necessarily involve the investigation of the construct validity of scores produced by the two tests. And construct validation in turn must begin with hypotheses about both the abilities the tests measure and the relative effects of test methods on test performance.

The ability and task analyses described here are being conducted as part of a project aimed at examining the comparability of two distinct batteries of English as a foreign language proficiency tests: the Cambridge Certificate of Proficiency in English (CPE) and First Certificate of English (FCE) on the one hand, and the Test of English as a Foreign Language (TOEFL), the Speaking English Assessment Kit (SPEAK) and the Test of Written English (TWE) on the other. The results of these analyses will be a set of hypotheses about similarities and differences in the language abilities measured by the various tests in these batteries and about similarities and differences in the test methods employed.

These analyses are guided by the theoretical frameworks of communicative language ability and test method facets described by Bachman (forthcoming). According to this description, communicative language ability consists of language competence, strategic competence and the psychophysiological mechanisms that are involved in language use. Test methods are described in terms of the following sets of facets: testing environment, instructions, input, response, and relationship between input and response.

Examples of item and parts of tests from both the CPE and the TOEFL will be used to illustrate the application of these theoretical frameworks to examining the comparability of language abilities and test methods across tests. This examination is pertinent to the assessment of content relevance and coverage of these tests. It is also useful for generating hypotheses that will guide the analysis of observed interrelationships among item and test scores on these different tests, which is relevant to the assessment of construct validity.

The use of these frameworks for analyzing test tasks reveals the complexity of potential interactions among test input and response facets and the abilities that a given task measures. It also reveals potential mismatches between the abilities measured and the abilities that may be involved in responding to test tasks. Implications of these analyses for the refinement of the theoretical frameworks are discussed, along with directions for further research into the way test takers process and respond to test tasks.
THE CONSTRUCT VALIDATION OF SELF-RATINGS OF COMMUNICATIVE LANGUAGE ABILITY

Lyle Bachman
University of Illinois at Urbana-Champaign

Adrian Palmer
University of Utah

(See Symposium on Self-Assessment)

THE CONTENT VALIDITY OF SOME ORAL INTERVIEW PROCEDURES:
AN ANALYSIS OF COMMUNICATION PROBLEMS AND STRATEGIES

Michael Canale
Ontario Institute for Studies in Education

This presentation has two goals: (1) to propose a framework for analyzing oral communication and oral communicative language ability in a first or second language; and (2) to compare sample analyses of some major oral assessment techniques with sample analyses of several extended oral interactions in nonassessment situations. The oral assessment techniques are drawn from oral interviews conducted by trained interviewers and videotaped by ACTFL, the Interagency Language Roundtable, the Ontario Test of English as a Second Language project, and the Israeli Oral Bagrut Exam project. The extended oral interactions represent both small group discussions and one-on-one conversations involving adolescents and adults. The focus of analysis will be the extent to which various levels of information, problems and strategies characteristic of natural oral communication are represented in the content of major oral interviews.

The proposed framework for analysis assumes that oral communication involves at least four distinguishable but interdependent levels of work or attention, any of which can come into and out of prominence at any point during a communicative event (such as conversation). Briefly, these levels and their respective areas of problems and strategies are: cognitive—representation and adequacy of factual, conceptual, experiential and procedural knowledge; interactional—interpersonal face, cooperation and control; affective—involvment, sociocultural identity and personal style; and linguistic—breadth, automaticity, and accuracy (or naturalness) of linguistic forms and their interpretation.

With respect to oral interview procedures, this framework and the sample comparative analyses are of interest for content validity purposes in several ways. For example, current oral proficiency scales are more product than process oriented in their characterizations of language proficiency; they focus almost exclusively on convention-governed aspects of communicative language ability (e.g. pronunciation, sociocultural features, grammar) than on those nonconvention-governed aspects involved in "playing interaction by ear" or making sense of the unconventional; they are more instrumentally or task-performance oriented than they are oriented toward problems and strategies in cross-cultural communication; and they are more useful for making vertical distinctions on a measurement scale (e.g. from 0 to 5) than they are for describing horizontally—within a given level of proficiency—the breadth of characteristic processes, problems and strategies. While current oral interview procedures may be adequate for the purposes they were developed to serve, it is useful, and at any rate necessary, to understand better their adequacy as valid representations of communication and communicative language ability in general.
Linguistic Analysis of Test Passages and Items As a Method of Determining Construct Validity.

Aaron S. Carton
State University of New York at Stony Brook

Introduction and rationale. Construct validity of a test is mainly a logical concept. As empirical validity indices are established, they contribute to the determination of construct validity. However, there is also a logical step to be performed in examining the construct validity of tests that are used as the operational definitions of a concept. This study is an attempt to conduct a logical linguistic content analysis of items used in a French Reading Proficiency Test developed for ACTFL. The content analysis is based on the linguistic, semantic and pragmatic dimensions specified in an earlier paper by Kaya-Carton and Carton (1986). The logical analysis is intended to relate the linguistic characteristics of the item-content to the item types used and to the difficulty level of the item translated into proficiency level. It also seeks to establish the correspondence between the logical analysis and actual findings concerning difficulty levels and dimensions measured by the items.

Procedure. The data base for the study is a 600 item test of French reading proficiency developed for ACTFL. The test was intended to cover six levels of proficiency. The items were based upon samples of French texts obtained from current written communications. Items were developed to include a variety of types and difficulty levels for each passage. The table of specifications which identified the linguistic, semantic and pragmatic dimensions had been used in constructing the items. This table also served as the basis for the content analysis to assure content validity. The logical validity of the categories derives from the agreement among a number of expert language educators.

Items were sorted by their types (e.g., multiple-choice, cloze, etc.) and cross-tabulated according to the dimensions provided in the table of specifications. The items were then arranged according to the level of proficiency they were presumed to measure. The presumed level of proficiency was chosen to retain the independence of the logical analysis from a concurrently conducted empirical study. Where an item measured more than one dimension, it was cross referenced to the other dimension(s) and an index of complexity was established. Thus each item had two indices: complexity and proficiency level. The indices determined through the logical analysis were then compared with the actual calibration of the items and dimensions obtained through the empirical study.

Hypotheses. The study sought to test several hypotheses. First, the logical analysis will yield the dimensions specified in the original table but the empirical analysis will collapse some of those dimensions. Second, the presumed proficiency levels of the items will correspond to the calibrated levels if the test-taking population is adequately sampled, and if the levels have empirical validity. Third, there will be a positive correlation between the logically derived complexity indices and the empirically determined difficulty calibrations of the items.

Findings. The first and third hypotheses were confirmed by the study. The paper will discuss possible explanations and implications for validating language proficiency tests.
VALIDATION OF A TAPE-MEDIATED, ACTFL/ILR-SCALE BASED TEST
OF CHINESE SPEAKING PROFICIENCY

John L. D. Clark
Defense Language Institute

At the 1986 Testing Colloquium, the author presented a paper describing the development, in four alternate forms, of a "semi-direct" (tape- and booklet-mediated) test of Chinese speaking proficiency. Subsequently, a validation study of the test was conducted, in which 32 native English-speaking learners of Chinese were each administered, under a Latin square design, two of the four test forms, together with a face-to-face speaking proficiency interview conducted by two trained ACTFL testers; these testers also served as scorers for the semi-direct test forms. Pearson product-moment correlations of .88 to .98 were obtained between the semi-direct tests and the live interview, suggesting a high level of criterion-related validity for the semi-direct approach as a potential alternative to live interviewing in situations where the latter procedure is not economically or administratively feasible.

Interrater reliabilities for individual forms of the semi-direct test ranged from .89 to .93. Test-retest reliabilities for various permutations of test form and scorer across both testing occasions ranged from .90 to .99. Notwithstanding the observed high correlations for both interrater and test-retest analyses, crosstabulations of the actual level scores assigned by the two raters revealed appreciably more "generous" scoring behavior on the part of one rater, for both live and semi-direct testing modes. A strong practical implication is that training and quality control of rater performance in operational testing situations should include close examination of the absolute values of the scores assigned as well as their linear relationships.

A feedback questionnaire administered on completion of testing indicated that most examinees considered their level of speaking ability to have been adequately probed under both live-interview and semi-direct conditions. However, the semi-direct tests were considered more difficult and more anxiety-producing than the live interview, as well as posing more "unfair" questions than the face-to-face test. To the question "assuming that you would receive the same score through both techniques, would you personally rather take a live interview or taped test in order to show your speaking ability?", 89 percent of the respondents favored the live interview.

Principal conclusions suggested by this study are that (1) high linear correlations are obtainable between direct ACTFL/ILR interview-based tests and tests designed to provide similar proficiency assessments through semi-direct means; (2) close attention should be directed to the absolute values of scores assigned across raters, both for score interpretation and for initial rater training and quality control; and (3) notwithstanding essentially similar scoring results for direct and semi-direct testing modes, examinee preference as to testing approach quite clearly favors the live, interactive procedure.
AN EXPLORATORY MODELING SURVEY OF THE TRAIT STRUCTURES OF EXISTING COMMUNICATIVE LANGUAGE TEST DATASETS.

Fred Davidson
University of California at Los Angeles

Recent developments in language testing have centered on implementation of multifaceted communicative language ability models. In fact, at some language testing sites, multifaceted communicative tests are already being developed to make workaday decisions about student proficiency. This trend stands in contrast to older "unifaceted" language tests which viewed language ability as governed by a single mental trait (Oller, 1979, 1983; Part I). This older, nondivisible model of communicative language ability has been pretty well laid to rest (Bachman and Palmer, 1982; Fouly, 1985).

Theorists have proposed several varieties of multifaceted communicative models (Canale and Swain, 1980; Swain, 1986; Bachman and Clark, 1987). There appears to be agreement that such a model will include the older nondivisible language trait as a single facet, while adding other facets such as control of register, ability to process discourse organization, facility with strategic conversational maintenance, and so on.

What remains to be done is twofold: first, language testing needs a series of confirmatory modeling studies of the theorists' various proposed models. Second, an exploratory modeling survey is needed on the existing (and therefore influential) communicative tests to determine what models seem to be operating at present in field use.

The present study undertakes the second task. It presents an exploratory modeling survey of ten language test datasets. Eight communicatively-designed test datasets are modeled: from the U.S., England, Canada and Sri Lanka. In addition, two supposedly undimensional "stock" English placement datasets from UCLA are modeled as controls. Modeling is performed using the recently developed TESTFACT program (Wilson, Wood, and Gibbons, 1986). This program provides a full information maximum likelihood factor analysis of smoothed interitem tetrachoric coefficients; thus, TESTFACT is in step with the very latest developments in item-level factoring. For each dataset, where necessary (due to computational or interpretational problems), the smoothed matrix produced by TESTFACT is modeled by other methods: a guessing-corrected unweighted least-squares factor extraction, multidimensional scaling, and higher-order factor extraction.

The study concludes with a discussion of the various trait structures uncovered in each dataset. Factors (and dimensions in the MDS analyses) are labeled with reference to prevailing theoretical multifaceted communicative models, and with reference to advice from the developers of each test. The study also compares the datasets with respect to (1) the number and nature of the traits uncovered, (2) intertrait correlations, and (3) possible higher level factor structures of communicative language ability. If an available theoretical model best fits a dataset analyzed, then this study could be said to help validate that model.
A STUDY OF VALIDITY CHARACTERISTICS OF THE SPEAK TEST

Dan Douglas
Iowa State University

Spencer Swinton
Educational Testing Service

Researchers at seven institutions (UCLA, Oregon, Iowa State, Minnesota, Ohio State, Penn State and ETS) are collaborating on a study of the SPEAK test, particularly its use in assessing the spoken English proficiency of international candidates for teaching assistantships. Among the questions being investigated are whether local scoring may differ in reliability from standard scoring as conducted at ETS in the Test of Spoken English program; whether, even if comparable in reliability, the SPEAK scorers may reliably employ higher or lower score thresholds, yielding scores systematically different from TSE scores; and whether more emphasis may be placed on certain dimensions (e.g., pronunciation) in assessing overall comprehensibility, yielding a measure of a slightly different construct than that assessed by a standard TSE scoring, a construct that may differ in validity from that of TSE scores. In addition to these central issues of reliability, score comparability and general construct validity, questions are being addressed concerning differential validity across institutions and across departments within institutions, form comparability, the use of SPEAK as an outcome measure in remedial courses, and the use of locally-developed measures, such as structured interviews or video-taped micro-teaching tasks, to supplement SPEAK scores.

PROCEDURE: At each of the six universities, SPEAK tapes are being double-scored, in a counterbalanced design tailored to the institution, so that each rater is paired with each other rater on the same number of tapes. Approximately 27 SPEAK tapes are being rated by each rater. The 27 tapes include six calibration tapes, previously scored at ETS, to be scored by raters at all six institutions. The approximately 650 ratings (27 tapes x 4 raters x 6 sites) are to be sent to ETS for analysis. In addition to interrater reliabilities, the mean and variance of ratings of calibration tapes by rater, institution, and overall institutions will be computed. The imputation of ratings by means of the EM algorithm will be explored as a means of adjusting for rater differences.

For those candidates actually assigned as teaching assistants, student ratings of instruction, supplemented by language-specific ratings, will be administered at the end of the first semester to one class taught by the international TA and to one control class in the same department taught by a native English-speaking TA. These data, along with SPEAK and TOEFL scores and background data from a questionnaire administered to the international TAs, will form the basis of a validity study which will take into account institution, department type, and background characteristics. Corollary studies of locally-developed measures will differ on each campus, but all will link SPEAK scores to the local measures.

The proposed colloquium presentation will be a preliminary report on data collected and analyzed to date. The discussion will be directed toward considering further analytical techniques, refinement of the data, and possible conclusions and outcomes of the research.

COLLABORATIVE ASSESSMENT BY MASTER'S CANDIDATES IN A TUTOR-BASED SYSTEM

Leslie Dickinson
Moral House College of Education

Gillies Haughton
The Scottish Centre For Education Overseas

(See Symposium on Self-Assessment)
MULTIPLE-CHOICE CLOZE ITEMS AND THE TOEFL

Gordon A. Hale, Donald A. Rock, Marilyn M. Hicks
Educational Testing Service

Charles Stansfield
Center for Applied Linguistics

Frances B. Hlinotis

John W. Oller, Jr.
University of New Mexico

Multiple-choice (MC) cloze items were developed and classified according to a scheme consisting of four categories that involved grammar, vocabulary, and reading comprehension in varying degrees. The objective was to determine if the different categories of MC cloze items related differentially to the various parts of the TOEFL, thus providing evidence of discriminant and convergent validity of these categories. The MC cloze items were included in a regular operational administration of the TOEFL. Analyses were conducted for each of the nine most heavily represented language groups, with a total of 11,290 subjects.

Exploratory and confirmatory factor analyses for the basic TOEFL were performed first to provide a basis for relating the MC cloze items to the TOEFL structure. These analyses suggested that, from a practical standpoint, TOEFL performance can be adequately described by just two factors, which relate to (a) Listening Comprehension (TOEFL Section 1), and (b) the other four parts of the test -- Structure and Written Expression (the two parts of TOEFL Section 2), and Vocabulary and Reading Comprehension (the two parts of Section 3).

Examinations of the MC cloze items showed that the correlations among scores for the four MC cloze categories were about as high as their reliabilities, thus providing no strong empirical evidence that the four categories of cloze items reflected distinct aspects of English proficiency.

Correlational analyses related the four MC cloze categories to the five parts of the TOEFL listed above. As expected, the correlations with Listening Comprehension were substantially lower than the correlations with the other four parts. There was a slight tendency for MC cloze items that involved a combination of grammar and reading to relate more highly to the Structure and Written Expression parts of the TOEFL than to the Vocabulary and Reading Comprehension parts, while the reverse was true for MC cloze items involving a combination of vocabulary and reading. This pattern was observed across language groups and was thus a relatively consistent finding. However, the differences among correlations did not appear substantial enough to be of practical significance. Multiple regression analyses were performed, using total MC cloze score as the criterion and the parts of the TOEFL as predictors. The resulting multiple Rs were mostly in the lower to upper 90s, suggesting a high degree of overlap in the skills being tested by the MC cloze test and the TOEFL.

It is perhaps not surprising that the relations between MC cloze categories and parts of the TOEFL did not differ substantially. Given that neither the internal analysis of the TOEFL nor the internal analysis of the cloze test indicated measurement of distinct non-aural skills, differential relations between the various MC cloze categories and the various non-aural parts of the TOEFL would not be expected. The data thus conform with the view that skills associated with grammar, vocabulary, and reading comprehension are interrelated.

Certain unresolved issues could benefit from examination in further research. For example, it would be useful to explore other schemes for classification of MC cloze items. Also, it would be of value to study the relation of various categories of MC cloze items to external criterion measures. Investigation of issues such as these would help determine whether it is possible to identify MC categories that tap separate aspects of English proficiency.
CONSTRUCT VALIDATION OF COMMUNICATIVE WRITING PROFILES

Liz Hamp-Lyons
University of Michigan

Grant Henning
Gerald De Mauro
Educational Testing Service

Commensurate with current developments in the testing of writing, the ongoing articulation of communicative testing theory, and the emerging of powerful new psychometric methodology, it seemed appropriate to the authors to investigate the validity of certain communicative writing profiles for use with writing tests such as the Test of Written English (TWE).

Accordingly, we propose to gather approximately 100, 30-minute TWE writing samples and approximately 100, 50-minute writing samples of a university writing achievement test. Using trained multiple independent raters, we plan to apply the New Profile Scale (Hamp-Lyons, 1987) in the marking of all writing samples. In the analysis, we propose to employ both multitrait-multimethod validation procedure (Campbell and Fiske, 1959) and, following the testing of appropriate assumptions, Rasch Model Partial Credit Analysis (Wright and Masters, 1982). In this way we hope (1) to gain insight into the convergent and discriminant validities of the constructs of communicative quality, organization, argumentation, linguistic appropriacy, and linguistic accuracy, and (2) to better understand the scalar properties of these constructs including both the nature of the score dispersion along the latent continuum they define and the associated degree of fit to the psychometric model on the part of the writing samples, the scalar points and the measurement constructs, and (3) to observe the comparative results of applying the statistical methodologies in the analysis of writing samples of different length and style.

It is understood that the New Profile Scale (NPS) was not explicitly developed for use with the two kinds of writing samples considered in this study, nor is its use being advocated over other scales that may already be in use. Nevertheless, for research purposes it is considered useful to investigate the extent to which any or all of the constructs employed in the NPS might be generalizable for application in other evaluative contexts.
THE RELATIONSHIP BETWEEN CANDIDATES’ SUBJECT SPECIALIZATION AND THEIR PERFORMANCE IN A TEST OF ENGLISH FOR SPECIFIC PURPOSES (ESP)

Peter Hargreaves
The British Council

John Foulkes
University of Cambridge

The ELTS Test is an ESP test which assesses candidates' communicative language ability both in general contexts and in contexts related to their broad subject areas. It is made up of a General Section (testing general proficiency and listening) and a Modular Section (testing study skills, writing and speaking), which is available in six broad subject areas: Life Sciences, Physical Sciences, Technology, Medicine, Social Studies and General Academic.

A particular concern for the construct validation of an ESP test like ELTS, which includes subject-specific modular options, is the question of degree of specificity: how many and how specific do the modular options need to be in order to capture efficiently systematic variation in the target population? What evidence is there to support the current six modules rather than, say, ten or three? Various aspects of this question have been investigated (e.g., Alderson and Urquhart 1984, and Criper and Davies forthcoming). The present research focuses on the following question: Does the degree of fit between a candidate's subject specialization and the module s/he is entered for have a significant effect on his/her performance in that module?

The research is based on the results of 1,000 ELTS candidates and their Test Report Forms (TRFs), which provide information about candidates' subject specializations. Since these specializations cover a very wide range and there are only six ELTS modular options to choose from, it is inevitable that some specializations fit more readily than others into a given ELTS module. Particular difficulties are encountered with subjects which intersect modular areas, e.g., Public Health (Medicine and Social Studies) or are interdisciplinary, e.g., Food Engineering (Life Sciences and Technology). For a variety of reasons some candidates may also be entered for a module inappropriate to their subject specialization.

In the case of the 1,000 candidates studied, it was hypothesized that those whose subject specialization closely matched the module for which they were entered should achieve higher scores in the modular subtests than candidates of equivalent general language ability, whose subject matched their module less well or not at all. According to this hypothesis, a student doctor, for example, entered for the Medicine module should score higher on the modular component than a student of Public Health or a student of Library and Information Studies (actual example) also entered for the Medicine module (provided they were of equivalent general language ability).

Using their TRFs, the 1,000 candidates were first categorized as either 'close fit', 'loose fit' or 'non-fit' on the basis of judgements about their subject specialization in relation to module taken. The three categories of candidates were then matched according to their scores in the General Section of the test. For each module the matched candidates were then compared with respect to their performance on the modular subtests.

The analysis of the data is currently being carried out and will be reported at the colloquium. The implications for the construct validation of ELTS and ESP test design will be explored.
THE VALIDITY OF SELF-ASSESSMENT TECHNIQUES IN LANGUAGE LEARNING:
COMPARING ABILITY AND SELF-ASSESSMENT OF ADULT EGYPTIAN EFL STUDENTS:

Russanne Hozayin
American University in Cairo

(See Symposium on Self-Assessment)

THE RELATIONSHIP OF SELF-ASSESSMENT TO EXTERNAL ASSESSMENT MEASURES OF
DUTCH AS A SECOND LANGUAGE

Anne-Marie Janssen-van Deiten
Katholieke Universiteit

(See Symposium on Self-Assessment)
EMPIRICAL COMPARISON OF THREE ITEM-CALIBRATION METHODS IN VALIDATING FRENCH READING PROFICIENCY LEVELS

Esin Kaya-Carton
Hofstra University

Introduction and rationale for the study. When reading proficiency tests are developed, passages of text and items that accompany them are selected on the basis of a priori judgements of the levels for which these passages and items are appropriate. Item calibrations are empirically determined mainly to validate these judgements. However, empirical calibration of various item types may also be viewed as a way to validate the proficiency levels themselves. Since the calibration method may have some effect on the empirical findings, it is very important to investigate whether different methods of statistical analysis yield different calibrations, and which method, if any, may best approximate the proficiency levels accepted by ILR and ACTFL.

Description of the study. Three methods of item calibration were selected for empirical comparison of their results: (a) the two parameter unidimensional Rasch analysis, (b) the two parameter multidimensional Rasch analysis, and (c) the Boolean factor analysis of item intercorrelations. The data for these analyses were obtained in the context of an ACTFL project to develop a computer interactive French Reading Proficiency Test. Data were collected using 600 items based on 54 reading passages, developed to measure six levels of reading proficiency. Approximately 400 subjects took each level of the test, totalling about 2,400.

The possible error in assuming that the reading proficiency is unidimensional was already discussed in a previous paper (Kaya-Carton and Carton, 1985). However, whether different methods of calibration in fact yield different results had yet to be tested. Also, since the multidimensional Rasch analysis is a relatively new concept, and so far does not seem to have been applied to language proficiency testing, it seemed prudent to conduct a methodological comparison to determine if it yielded calibrations that were more valid than calibrations obtained by the unidimensional method.

Factor analysis is another method to determine multidimensionality. It is possible to use item factor loadings to calibrate items on given dimensions. However, although such calibration may yield a Guttman scale type of ordering, it would be difficult to attribute such ordering to item difficulty per se without comparing the factor analysis results with a method that is designed specifically to yield item-difficulty indices. Thus, logically, one would tend to favor the use of the multidimensional Rasch analysis as the best method of approach. Yet, given that this method is relatively untested, and somewhat more cumbersome than the other methods, one needs to determine empirically whether it is, in fact, the best method to utilize.

Hypotheses. Four hypotheses were tested in the study: 1) the unidimensional Rasch analysis will tend to have greater error in the item calibrations than the multidimensional Rasch analysis; 2) the Boolean factor analysis will yield the same number of dimensions as the multidimensional Rasch analysis, but will yield different values for item calibration on each of the dimensions; 3) the multidimensional Rasch analysis will correspond the best to the proficiency levels represented by the test items; 4) all three methods will yield the same number of proficiency levels.

Results. The results obtained show that some of these hypotheses were confirmed but others were not. Although the results tend to support the validity of the proficiency levels there are differences among the levels with regard to the clarity of the calibrations. The multidimensional Rasch analysis provided the most readily interpretable results. On the other hand, the comparisons raise some doubt on the validity of the upper proficiency levels with the samples used in the study. A number of methodological issues also appeared in the empirical comparisons. The paper will present the actual data and discuss the implications for the validity of the tests as well as the validity of the proficiency levels themselves.
PERSON DIMENSIONALITY IN LANGUAGE TEST VALIDATION

Brian Lynch
Fred Davidson
University of California at Los Angeles

Grant Henning
Educational Testing Service

The validation of tests of communicative language ability should consider elements related to test-takers as well as the tests themselves. Elements such as native language background, residency status, and length of stay in the target language culture may well provide important information in test validation efforts, inasmuch as such elements may define statistically significant examinee dimensions that would result in unpredictable variability of test item response statistics.

The present study will investigate person dimensionality in relation to students taking the UCLA English as a Second Language Placement Examination (ESLPE). Test data from the Fall 1987 administration of the ESLPE (n=850) will be analyzed using cluster analysis to determine if statistically significant examinee subgroups exist. An attempt will then be made to label such subgroups with reference to student demographic information such as the elements mentioned above. Next, the responses of these groups will be factor analyzed to determine if there are differences in trait structure across the groups. Finally, the relation of these subgroupings to item response statistics will be examined--specifically, unpredictable change in Rasch calibrations and fit statistics.

If no statistically significant subgroups are found to exist, or if the subgroups resist logical labelling, then explanations for this finding will be explored. In such a case, the appropriacy of cluster analysis in the determination of person dimensionality will be discussed. If statistically significant subgroups are found to exist and can be logically labelled but their test responses display similar factor analyses and Rasch statistics, then this will be taken as evidence for test validity.

In general, this study is concerned with determining the degree to which person dimensionality and its effect on test item response patterns can provide useful information in the validation of communicative language tests. Statistically significant subgroups, if their responses display a similar trait structure (via factor analysis) represent evidence of a kind of "global" validity. Further, if they are similar in their Rasch item calibrations, they then display validity at the item level, or "local" validity. The integration of cluster analysis, factor analysis, and Rasch modeling, in this way, can provide valuable assistance to test validation.
HOW NATIVE LANGUAGE AND LEVEL OF ENGLISH PROFICIENCY
AFFECT THE STRUCTURE OF THE
TEST OF ENGLISH AS A SECOND LANGUAGE (TOEFL)

Philip Oltman
Lawrence J. Stricker
Educational Testing Service

An individual's performance on the TOEFL may reflect the influence of both their native language and their level of English proficiency. The aim of this study was to appraise the joint effects of native language and English proficiency on the structure of the test. We wanted to explore how the items clustered together, and whether the nature of the clusters differed across groups defined on the basis of their native languages and their overall performance levels on the TOEFL. The interrelations among TOEFL items, using all of the information provided by the various responses to the items (the four alternatives, omitted, and not reached), were analyzed by three-way multidimensional scaling for samples of examinees systematically varying in native language and level of English proficiency. Four dimensions were identified: three corresponded to the sections of the test, and the fourth was an end-of-test phenomenon. We then applied a hierarchical cluster analysis to the multidimensional scaling results to facilitate interpretation of the dimensions. The dimensions were predominantly defined by easy items and were most salient for low-scoring examinees. The salience of the dimensions did not differ for the various language groups, except for the end-of-test dimension. We conclude that the results support the TOEFL's construct validity, in that the dimensions and clusters we found correspond closely to the three sections of the test. We also conclude that the test's interpretation must vary with examinees' English proficiency, that easy and difficult items differ in their potential for diagnosis and global screening, and that the dimensionality of the TOEFL and of competence in English depends on the level of examinees' English proficiency.

SELF-ASSESSMENT OF LANGUAGE PROFICIENCY: RATIONALE AND APPLICATIONS

Mats Oskarsson
Gothenburg University

(See Symposium on Self-Assessment)
CONTENT AND CONSTRUCT VALIDATION OF LISTENING COMPREHENSION TESTS

Elana Shohamy
Ofra Inbar
Tel Aviv University

This paper reports on a research study which examined the effect of text and question types on scores which students obtain on listening comprehension tests. EFL listening comprehension tests based on two types of texts each representing three text types—a news broadcast, a lecture and a consultative dialogue, were administered to 150 high school students. The passages were controlled for information and lexis but varied in the degree of oral features they contained. Subjects answered identical questions to enable comparison of performance on the three different text types and the questions were classified into global and local question types according to strategies used for text processing. Results showed that listening comprehension scores are significantly affected by choice of text type, and that performance related to local cues in the texts yields significantly higher scores than performance related to global elements. The implications of these results for both content and construct validity of listening tests will be discussed with respect to the types of texts and types of questions that should be used on listening tests and to the text types which best represent the knowledge of this trait.

FACILITATING ACCURATE SELF-ASSESSMENT OF FUNCTIONAL LANGUAGE SKILLS

Bernard Spolsky
Bar Ilan University

(See Symposium on Self-Assessment)
A LONG TERM RESEARCH AGENDA FOR THE TEST OF WRITTEN ENGLISH

Charles Stansfield
Center for Applied Linguistics

Jacqueline Ross
Educational Testing Service

The Test of Written English (TWE) became an operational part of the TOEFL program in July, 1986. During its first year, over 100,000 examinees took the test at three administrations. Interest in the test among university admissions officers has been strong, and many institutions are considering the policy of requesting TWE scores of applicants. Others, who have not yet decided to require the TWE, are receiving TWE score reports from applicants. This has generated many requests for information about the reliability and validity of the test. Other groups are interested in it as well. These include ESL teachers who may be using TWE scores to place students within the writing instruction program of an English language institute, specialists in second language testing, and specialists in composition. The TOEFL Policy Council, the TOEFL Research Committee, the TOEFL Committee of Examiners, the TWE Core Readers, and TOEFL program staff at large, all agree that while the studies that describe the rationale and development of the TWE (Bridgeman and Carlson 1983, Carlson et al. 1985) provide considerable information, much additional research on the test is needed. However, the nature of this research is so diverse and the number of issues to be examined is so large and conceptually problematical that a systematic overview is desirable. For that reason, in April 1987 the TOEFL Research Committee commissioned the development of a long-range research agenda for the TWE. This paper is the draft final report for that project.

The long-range research agenda was developed with input from numerous individuals and groups. A special meeting was held at the ETS campus in Princeton, NJ on September 12, 1987. Attending the meeting were representatives of the TOEFL Research Committee, the TOEFL Committee of Examiners, and the TWE Core Readers. ETS designated appropriate representatives from the Research, Statistical Analysis, Test Development and Program Direction areas. The meeting generated many fruitful ideas for research.

One week later at the regular meeting of the TWE Core Readers, the group devoted two hours to a discussion of a long-range research agenda. This discussion produced additional ideas. Two hours or more were spent discussing the TWE research agenda at the October 1987 meeting of the TOEFL Research Committee and the TOEFL Committee of Examiners. Additional input was provided at these meetings which was incorporated into a draft of the paper. Subsequently, a draft of the paper was distributed for review by participants in the September 12 special meeting and to other ETS staff. Following input from participants in the Tenth Annual Colloquium on Language Testing Research, the report will be submitted to the TOEFL Research Committee in March and to the TOEFL Policy Council in May 1988.
A SYMPOSIUM ON SELF-ASSESSMENT TECHNIQUES IN LANGUAGE LEARNING:
IMPLEMENTATION, VALIDITY AND RATIONALE

Organizer:
Russanne Hozayin
American University in Cairo

While the most common way to assess language ability has been the traditional test format, another technique being used by those who deal with adult second language learners is to ask the learners to assess their language skills. This mode of "asking" typically includes a very detailed set of questions which are based on a thorough analysis of the types of situations in which the students will use the target language.

Although several persons have been using self-assessment techniques in lieu of or in addition to traditional tests (e.g., Oskarsson, and von Elek in Sweden, Le Blanc in Canada, Raasch in Germany, and Bachman and Palmer in the U.S.), no recent symposium has specifically focussed on the issues connected with self-assessment. Chief among these issues are the rationale for using self-assessment in language learning, the validity of this technique, and its relationship to other types of language ability measures. This symposium will not only serve to focus attention on self-assessment as an alternative technique to traditional testing, it will also help bring together several of the leading persons in the field of language testing research and those who have been working in the self-assessment field, in order to enhance the work being done by both groups.

THE CONSTRUCT VALIDATION OF SELF-RATINGS OF
COMMUNICATIVE LANGUAGE ABILITY

Lyle Bachman
University of Illinois at Urbana-Champaign

Adrian S. Palmer
University of Utah

The trait structure of an experimental self-rating test of communicative language ability was investigated through the use of MTMM design and CFA procedure. The language abilities we attempted to measure were hypothesized to comprise three main traits: grammatical competence, pragmatic competence, and sociolinguistic competence. The subjects were 166 non-native English speakers from the Salt Lake City area. The results of this study indicate that self-ratings can be reliable and valid measure of communicative language abilities. The obtained reliabilities were much higher than had been expected, and all the self-rating measures had strong loadings on a general factor. In addition, some measures proved to be reasonably good indicators of specific language abilities. The measures of grammatical competence appear to be better indicators of this trait than are the measures of pragmatic and sociolinguistic competence. And of the three question types used, the most effective appears to be that which asked about subjects' perceived difficulty with various aspects of the language. The least effective question type was the so-called "can-do" question, which in general was the most affected by the test method.

24
COLLABORATIVE ASSESSMENT BY MASTER'S CANDIDATES IN A TUTOR-BASED SYSTEM

Leslie Dickinson
Moral House College of Education

Gillies Haughton
The Scottish Centre For Education Overseas

This paper will present the results of a scheme of collaborative assessment which has been running for two years with students on a master's level course in English language teaching. The method of self-assessment which is used is as follows. Having completed an assignment, the course member can grade herself or himself A to E using a set of criteria previously negotiated between the tutor and the student group. This grade is compared with the tutor's grade and if these grades are different the course member can negotiate with the tutor-on the basis of the criteria-towards a mutually acceptable grade. If the negotiation fails, then the work is passed to another tutor acting as a referee whose decision is final. The discussion will focus on the implementation of self-assessment in an actual program and the success of self-assessment compared to traditional methods of assessment, as seen by both course members and tutors.

THE VALIDITY OF SELF-ASSESSMENT TECHNIQUES IN LANGUAGE LEARNING: COMPARING ABILITY AND SELF-ASSESSMENT OF ADULT EGYPTIAN EFL STUDENTS:

Russanne Hozayin
American University in Cairo

The relationship between ability assessment and self-assessment techniques will be the focus of this paper, with results based on the self-assessment and ability assessment of approximately 1,500 adult Egyptian basic-level EFL students, both prior to and following a language training course. Problems that will be discussed include the differential self-assessment ability which characterized the students and setting the appropriate climate to gather valid self-assessment data.

THE RELATIONSHIP OF SELF-ASSESSMENT TO EXTERNAL ASSESSMENT MEASURES OF DUTCH AS A SECOND LANGUAGE

Anne-Marie Janssen-van Deiten
Katholieke Universiteit

Adult learners of Dutch as a second language were assessed using both criterion tests of speaking, reading, writing and listening ability and self-assessment techniques. The results of these two types of approaches were compared and are reported on in this paper.
SELF-ASSESSMENT OF LANGUAGE PROFICIENCY:  
RATIONALE AND APPLICATIONS

Mats Oskarsson  
Gothenberg University

1. Introduction  
2. Why self-assessment?  
3. Self-assessment techniques and materials  
   3.1 Progress cards and forms  
   3.2 Questionnaires, rating scales, check lists  
   3.3 Diaries and log books  
   3.4 Informal self-assessment  
   3.5 Video and audio cassettes  
   3.6 Computer-assisted assessment

FACILITATING ACCURATE SELF-ASSESSMENT OF  
FUNCTIONAL LANGUAGE SKILLS

Bernard Spolsky  
Bar Ilan University

While language tests generally require the cooperation of the person being tested, self-assessment assumes that the cooperation is honest and willing. Assuming that there is no obvious reason for the subject to cheat, the accuracy of self-report depends on the preciseness of the questions and their salience. The technique is thus suitable for assessing functional communicative language skills in a second language. A study is reported in which high levels of reliability and validity were achieved with self-report.