PROGRAM

SEVENTH ANNUAL LANGUAGE TESTING RESEARCH COLLOQUIUM

HENRY CHAUNCEY CONFERENCE CENTER
EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
APRIL 6 - 9, 1985
Dear Participant:

Welcome to the seventh annual Language Testing Research Colloquium. We hope you will enjoy your stay at the Henry Chauncey Conference Center, and that you will profit professionally from the Colloquium presentations and your visit to Educational Testing Service.

Meals at the Conference Center are available on a specified schedule. Breakfast and lunch are served buffet style from 7:30 to 9:00 a.m. and noon to 1:00 p.m., respectively. The Conference Center contains an exercise room and a sauna on the bottom floor of the west wing. A bar, located directly above in the West Reception Room, is open daily from 4:45 to 11:00 p.m. You are also invited to walk or jog on the ETS grounds. For joggers, a map of cross country trails on ETS property can be obtained at the registration desk.

The Seminar on Item Response Theory scheduled for Saturday, April 6, will be taught by Martha Stocking of the ETS Research Division. Dr. Stocking works with Fred Lord, whose research provides much of the theoretical basis of IRT. Saturday evening, there will be a commercial demonstration of Visi-Pitch, a machine that measures the prosodic features of speech.

The traditional presentation of papers at the Colloquium will take place on Sunday and Monday, April 7 and 8. A banquet is scheduled for Sunday evening. Information on dining and entertainment options for Monday evening will be covered in the Announcements on Monday morning. Following the Colloquium on Tuesday, April 9, you will have a choice of activities that will allow you to learn more about ETS.

An abstract of each Colloquium presentation is included herein. Since there is only a limited amount of time for each presentation, your reading the abstracts beforehand will enable the presenters to progress more easily through the material.

If you have any questions about ETS, do not hesitate to ask me or one of the ETS staff attending the Colloquium. Questions about the Conference Center should be directed to Conference Center staff. We hope you will enjoy your visit.

Sincerely,

Charles Stansfield
Colloquium Chairperson
Saturday, April 6

9:15 a.m.  BARN  Seminar  
"An Introduction to Item  
Response Theory".  .  .  .  .  .  .  .  .  .  Martha Stocking  
Educational Testing Service

10:30 a.m.  BARN  Break

10:45 a.m.  BARN  Seminar

12:00  HCCC  Lunch
      DINING  ROOM

1:00 p.m.  BARN  Seminar

3:00 p.m.  BARN  Break

3:15 p.m.  BARN  Seminar

4:45 p.m.  BARN  Adjourn

8:15 p.m.  BARN  Demonstration  
"Visi-Pitch: An Instrument  
to Evaluate Pronunciation."  .  .  .  John Crump, President  
Kay Elemetrics Corporation
### The Seventh Annual Language Testing Research Colloquium

#### PRESENTATION OF PAPERS

**Sunday, April 7**

<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Title</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:15 a.m.</td>
<td>BARN</td>
<td>Welcome</td>
<td>Charles Stansfield</td>
</tr>
<tr>
<td>8:30 a.m.</td>
<td>BARN</td>
<td>&quot;Computer Assisted Language Testing: A Cloze Investigation&quot;</td>
<td>Alan Davies, Clive Criper</td>
</tr>
<tr>
<td>9:30 a.m.</td>
<td>BARN</td>
<td>&quot;Can High Technology Help Us Test Communicative Competence?&quot;</td>
<td>Joyce Penfield</td>
</tr>
<tr>
<td>10:15 a.m.</td>
<td>BARN</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>10:30 a.m.</td>
<td>BARN</td>
<td>&quot;Learner-generated Distracters for MC Cloze: Item Difficulties and Person Ability Estimates&quot;</td>
<td>Virginia Straiff</td>
</tr>
<tr>
<td>11:15 a.m.</td>
<td>BARN</td>
<td>&quot;Computerized Adaptive Testing: Implications for Language Test Developers&quot;</td>
<td>Peter Tung</td>
</tr>
<tr>
<td>12:00</td>
<td>HCCC DINING ROOM</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:00 p.m.</td>
<td>BARN</td>
<td>&quot;An Evaluation of Several Methods of Routing and Branching Examinees to Levels of a Placement Test for the Test of English as a Foreign Language&quot;</td>
<td>Marilyn Hicks</td>
</tr>
<tr>
<td>1:45 p.m.</td>
<td>BARN</td>
<td>&quot;Item Banking Via DBase II&quot;</td>
<td>Grant Henning</td>
</tr>
<tr>
<td>2:30 p.m.</td>
<td>BARN</td>
<td>&quot;Development of a Tape Mediated ACTFL/ILR Scale Based Test of Chinese Speaking Proficiency&quot;</td>
<td>John L.D. Clark</td>
</tr>
<tr>
<td>3:15 p.m.</td>
<td>BARN</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>3:30 p.m.</td>
<td>BARN</td>
<td>&quot;Taxonomy of Language Attainment Scales&quot;</td>
<td>Pardee Lowe, Jr.</td>
</tr>
<tr>
<td>4:15 p.m.</td>
<td>BARN</td>
<td>&quot;Models of Second Language Competence: A Structural Equation Approach&quot;</td>
<td>Helmut Vollner</td>
</tr>
<tr>
<td>4:15 p.m.</td>
<td>WEST 5</td>
<td>&quot;Comparison of Integrated and Isolated Procedures for Testing Overall Oral Proficiency&quot;</td>
<td>Elana Shohamy</td>
</tr>
</tbody>
</table>
| 7:00 p.m.| HCCC DINING ROOM | Banquet                                                                                | "Some Reflections on Productive Language Testing" | Protase Woodford
# PRESENTATION OF PAPERS

**Monday, April 8**

<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Title</th>
<th>Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 a.m.</td>
<td>BARN</td>
<td>Announcements</td>
<td>Charles Stansfield</td>
</tr>
<tr>
<td>8:45 a.m.</td>
<td>BARN</td>
<td>&quot;The Promise and Threat of Computerized Adaptive Techniques in Assessing Reading Comprehension&quot;</td>
<td>Michael Canale</td>
</tr>
<tr>
<td>9:30 a.m.</td>
<td>BARN</td>
<td>&quot;Computerized Rasch Analysis of Item Bias in Selected ESL Subtests&quot;</td>
<td>Harold Madson</td>
</tr>
<tr>
<td>10:15 a.m.</td>
<td>BARN</td>
<td>Break</td>
<td>Jerry Larson</td>
</tr>
<tr>
<td>10:30 a.m.</td>
<td>BARN</td>
<td>&quot;Computer Adaptive Reading and Listening Tests for Foreign Languages: A Progress Report&quot;</td>
<td>June Phillips</td>
</tr>
<tr>
<td>11:15 a.m.</td>
<td>BARN</td>
<td>&quot;Item Selection from Pretests in Mixed Ability Groups&quot;</td>
<td>John H.A.L. de Jong</td>
</tr>
<tr>
<td>12:00</td>
<td>HCCC DINING ROOM</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>1:00 p.m.</td>
<td>BARN</td>
<td>&quot;Using Technology to Assess Second Language Proficiency through Cloze-tests&quot;</td>
<td>Winton Manning</td>
</tr>
<tr>
<td>1:45 p.m.</td>
<td>BARN</td>
<td>&quot;Computer Text Analysis in ESL Composition Teaching and Testing&quot;</td>
<td>Joy Reid</td>
</tr>
<tr>
<td>2:30 p.m.</td>
<td>BARN</td>
<td>&quot;The Correlation Between Machine and Human Ratings of TSE Reading Passages&quot;</td>
<td>Gary Molholt</td>
</tr>
<tr>
<td>3:15 p.m.</td>
<td>BARN</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>3:30 p.m.</td>
<td>BARN</td>
<td>&quot;The Use of Computers to Identify Two Different Stages of Language Acquisition for Thai Students&quot;</td>
<td>Kanchana Praphal</td>
</tr>
<tr>
<td>4:15 p.m.</td>
<td>BARN</td>
<td>Will Technology Improve Language Testing?: An Open Discussion</td>
<td>Julia Maceda-Willebrand</td>
</tr>
<tr>
<td>4:45 p.m.</td>
<td>BARN</td>
<td>Business Meeting</td>
<td>Charles Stansfield</td>
</tr>
<tr>
<td>5:00 p.m.</td>
<td>BARN</td>
<td>Au revoir</td>
<td></td>
</tr>
</tbody>
</table>
Tuesday, April 9

8:30 a.m.  J-Bldg.  Tour of ETS operations facilities
            Note: Vehicles will depart at this time from the
            Henry Chauncey Conference Center.

11:00 a.m.  P016  Test Development Procedures
            at ETS ..................... Protase Woodford

            P014  TOEFL, Pre-TOEFL, SLEP, and
            the TOEFL Essay Test ........ Russell Webster

12:00 noon  Lunch

1:00 p.m.   P016  Language Testing Research
            at ETS ..................... Gordon Hale

            P014  Test of Spoken English ........ Rodney Ballard

2:00 p.m.   P016  Test of English for
            International Communication ... Steve Stupak

            B016  ETS Test Collection .......... Alicia Dodd

            P012  History of ETS .............. Gary Saretzky
Computer Assisted Language Testing: A Cloze Investigation

In this paper we criticize the unthinking acceptance of testing as an obvious computer assisted language learning (CALL) application. Computers will not provide linguistic or testing insights that are not available in the parent disciplines. The two features of CALL that make it a potential asset to language testing are speed and memory. Speed enables the most appropriate item array to be presented for any given group of learners or any individual learner, or in criterion referencing for any given level of knowledge or skill. Memory provides for the closure of the gap between the population of all possible test items and the sample under test by drawing more and more learner specific samples. An example of a testing technique which a computer can carry out effectively but which cannot be easily done using paper and pencil is adaptive testing. This involves selecting items during the process of administering a test so that the items administered have the right difficulty level of the learner(s) under test. As well as discrete point tests, integrative tests can be used in this way. For example, cloze procedure can be used to make a test more efficient by increasing reliability by repeated selection of "items". Relevant cloze test data from current work at the University of Edinburgh are discussed.

Reference

Can High Technology Help Us Test Communicative Competence?

In the past few years a great deal of attention has been directed toward the use of language as a communicative tool and second language acquisition research and pedagogy now reflects an interest in naturalistic settings. Research in language proficiency testing is now concerned with "authenticity" of language. However, much of actual testing has had difficulty simulating a primary factor of authentic language, i.e., context. One specific type of high technology now exists which might be useful in capturing some crucial factors of authentic speech--interactive video. But first it will be necessary to analytically understand the nature of context and such factors that contribute to communicative competence.

This paper considers the various ways in which interactive-video through the use of discs and/or video-tapes might be helpful in testing various constructs of communicative competence. Initial discussion will be given to what should be tested i.e. what exactly accounts for "authenticity" of language and then various applications of these constructs to interactive video technology will be suggested.

NOTE:

*The author is in the process of working on a project to illustrate an application of interactive video to listening comprehension but this is only in preliminary stages. The video disc to be used is currently being pre-mastered and it may be possible to provide a demonstration at the conference if the project continues successfully.
Learner-generated Distractors for MC Cloze: 
Item Difficulties and Person Ability Estimates

Cloze procedure holds promise as a valuable tool in computerized-adaptive language testing (CALT), with multiple-choice (MC) cloze suggested as an alternative to open-ended cloze. MC cloze is potentially useful in circumstances where machine-scoring (and other analyses) is needed or preferred owing to speed, objectivity, and—in the case of CALT—increased accuracy of skill estimation. The value of learner-generated distractors for MC cloze was described in part at the 1984 Language Testing Colloquium, emphasizing traditional test statistics and introducing Rasch-calibrated item difficulties and person ability estimates obtained by hand via the PROX method (Wright and Stone, 1979). The 1984 data incorporated three forms of the Texas Agricultural and Mechanical University (TAMU) Reading Examination for international student placement.

Progress on the study (begun in 1983) will be reported. The current work includes Forms 1 through 4 of the TAMU Reading Examination with computer-analyzed results. The data were gathered from 1302 subjects who entered the university as freshmen, transfer, or graduate students. The test data were pooled for each Reading Examination form and mean square fit statistics analyzed for each test item.

The report will emphasize Rasch-calibrated item difficulties, person ability estimates, and mean square fit statistics of the four test forms. The results may be important for MC cloze development and also for further exploration of open-ended cloze in the CALT context.
Computerized Adaptive Testing: 
Implications for Language Test Developers

This paper discusses the implications of using an automated individualized-measurement procedure—computerized-adaptive testing (CAT)—to assess second language performance and demonstrates that the efficiency and applicability of CAT can be increased by incorporating some additional routines. Based on the item response theory, CAT is known to be more efficient and more accurate than conventional fixed-length tests employing multiple-choice items. It is suggested that the implications of CAT for language testing can be discerned by considering the following basic questions: What should be the number and characteristics of items included in an item pool for CAT? Which test item in the item pool should be administered first to a given individual? What should be the rules governing the selection of subsequent test items? When should a CAT session terminate? The advantages as well as the limitations of CAT for language testing may then be clarified. Several desirable optional features are illustrated by means of the results obtained with an interactive CAT program using simulated examinees and data banks. For example, a computerized adaptive test can be shortened if information about the examinees obtained prior to the test is used to select the first items administered to them. Optimal use of the items in an item bank and test security can be promoted by selecting the next item to be administered randomly from a few possible items instead of displaying the mathematically-determined "best" item. It is argued that CAT is a viable technology for language testing provided that its limitations are recognized and steps taken to adapt it to meet the special needs of language test developers.
An Evaluation of Several Methods of Routing and Branching Examinees to Levels of a Placement Test for the Test of English as a Foreign Language

In order to identify optimal methods of routing examinees to levels of a proposed computerized TOEFL placement test, simulation studies of several routing procedures were conducted using examinee responses on a form of TOEFL. The nature of most of the routing procedures considered in this study reflected their basic classificatory purpose as well as the restrictions of space on a floppy disk housing the entire testing system. Since it is expected that both computerized and paper-and-pencil tests will be developed, routing procedures appropriate for both testing modes were investigated. In general, the sequential testing methods were developed for use with the computerized tests, while fixed length, nonsequential methods were evaluated for use with the paper-and-pencil test versions.

This study focused on methods of testing using conventional item data, and revisited some of the problems associated with two-stage testing. The current study exploited the flexibility of computerized testing and employed variations in sequential testing at the routing level in order to improve the process.

Three basic methods of routing to second stage tests, and multilevel testing, a method of branching directly through the measurement test, were evaluated using actual responses on a form of TOEFL. The criterion for determining correct classification was the score obtained on the full length TOEFL. The methods included routing by a sequential administration of items, with the sequence depending on the correctness of the response to the previous set of items. This resulted in varying numbers of items presented to each examinee. A fixed length test equated to the TOEFL scale was evaluated for use with a paper-and-pencil version of the test. In addition, the sequential administration of items with decision rules based on an application of Bayes theorem was investigated. Finally, multilevel testing (a method in which the routing test is bypassed) was also simulated. All routing tests were limited to a maximum of ten items; the multilevel tests consisted of approximately one-half the items found in corresponding sections of the regular TOEFL.

The rates of misclassification for the routing tests were greatly decreased, in comparison with those observed in the literature, by defining overlapping test levels and by sequential administration of items. Fixed length, nonsequential procedures produced the highest rates of misclassifications as expected, and probably should be limited to tests with only two levels. The results for multilevel testing were excellent, resulting in a correlation of .95 between scores on the levels and regular TOEFL scores. This latter method seemed the most promising procedure for administering an adaptive test using conventional item data; however, some of the routing procedures utilizing sequential presentation of items may be useful for the rapid screening of large numbers of individuals if only a gross estimate of level of language proficiency is required.
Item Banking Via DBase II

A lot of interest in recent years has been focused on how best to develop and maintain test item banks for use with individual ESL programs. A merging of technological developments in microprocessing with measurement theory developments involving item response theory/latent trait measurement enables programs to maintain such item banks and to tailor tests to the existing program needs.

This paper reports on procedures for item banking and machine construction of language proficiency tests using DBASE II software and microprocessor hardware. Rasch model difficulty calibrations and a variety of other item statistics are stored along with content specifications and use data.

The advantages of this approach are:

1. Use is made of widely available software.
2. A variety of microprocessors may be used.
3. Large databases may be stored, manipulated, and retrieved (more than 65,000 items).
4. Alternate forms of tests may be easily constructed.
5. Test tailoring is facilitated.

The information presented will be based on development efforts with the ESL placement exam in use at the University of California, Los Angeles. A portable microprocessor will be provided for demonstration purposes. Discussion will center on specific problems to be overcome in the organization and maintenance of such an item bank. Information will be provided about the appropriateness of the measurement model employed for item bank construction and maintenance.
This paper describes the design, development, and initial administration of a tape-mediated "semi-direct" test of Chinese speaking proficiency for American and other English-speaking learners of Chinese. This test—which is designed to be scored and interpreted according to the ACTFL/ETS/ILR guidelines for direct proficiency tests, but also to be usable in situations where actual face-to-face testing is financially or administratively infeasible—makes use of several different elicitation techniques, including two-way "conversation"; picture-based narration; description of personal background and interests; and performance of a variety of situational tasks, including tasks requiring differing linguistic registers and affectively marked discourse.

Considerations in the selection of these particular question types by comparison to other possible formats will be discussed, as well as the perceived measurement advantages and disadvantages of tape-based administration of comparison to the live interview. Results of administration of a preliminary form of the test will be presented. The paper will conclude with a discussion of technical and practical aspects of test development for tape-based instruments, with recommendations for further research and validation activities in this area.
Taxonomy of Language Attainment Scales

Currently, several scales describe the functional levels attained by students or users of foreign language. A comparison of the scales, however, reveals several significant differences. This paper describes a taxonomy of language attainment scales, a taxonomy sensitive to these differences. Treated are the U.S. Government's Interagency Language Roundtable Scale (ILR), the American Council on the Teaching of Foreign Languages/Educational Testing Service's (ACTFL/ETS) Generic Proficiency Guidelines, the Australian Second Language Proficiency Rating Scale (ASLPR), the Canadian Public Service Commission (CPSC) Scales, and the British Council Scales. Topics discussed include: origin and commensurability, number of levels, distance between levels, number of skills (single or linked), uniformity in criteria or lack thereof, extent of focus on general or on specialized language, and finally, upward or downward branching; that is, the extent to which scales refer either to single or to multiple criteria. For example, the Australian scale derives from the ILR scale, yet proves incommensurate with the latter due to the replacement of the ILR's classic Educated Native Speaker (ENS) by more work-related or job-specific language users. This taxonomy arose through adapting the general-language ILR scale for describing linguistic performance for a special purpose, business.

Using the categories developed in the taxonomy, the paper's second section examines the suitability of such scales to research. This section results in part from Lowe's (1984) suggestion for combining the ILR scale with the derivative, commensurate ACTFL/ETS scale to conduct what is here termed macroresearch and in part by John L.D. Clark's reaction to that suggestion. The paper distinguishes two strands of second language acquisition research, microresearch customary in academia and macroresearch, more frequent in U.S. Government studies. While not a strict dichotomy, microresearch tends to emphasize detailed studies on small N's and is often of a longitudinal nature; while macroresearch may be defined as studies of large categories based on large N's. The former might typically examine such minute categories as Present Progressive, Copula, Aux (Prog.), Past Aux, etc.; while a typical example of the latter would look at the larger category of Grammar in relation to other categories such as Vocabulary, Sociolinguistics/Culture, etc.

A possible third section will examine a specific instance of macroresearch and discuss its description in terms of the taxonomy.
Elana Shohamy  
School of Education  
Tel Aviv University  
Tel Aviv 69978 Israel  

Comparison of "Integrated" and "Isolated" Procedures  
for Testing Overall Oral Proficiency  

With the increase in emphasis on communicative testing, oral-proficiency testing is receiving special attention—mainly through experimenting with different testing procedures. Most of these tests tap a single speech style and it is questionable whether they tap the whole domain of oral proficiency. The need to tap more than one speech style in the assessment of overall oral proficiency has been expressed by Clark (1980), Lowe (1980). Shohamy, Reves, Bejerano (1984) experimented with the development of an overall oral proficiency test which consisted of an oral interview, a role play, a reporting task and a group discussion. Comparing students' performance on these tests and on a single test leads the authors to call for the use of oral tests which consist of a number of speech styles in order to assess oral proficiency in a valid way. It is not clear, however, what the format of such a multiple speech-style test should look like. Is it necessary that the testing of these different speech styles be done by different testers on separate testing occasions, or perhaps these different speech styles can be tapped more efficiently by one tester in one testing event?

The paper that will be presented addresses this very question by comparing these two formats. It describes two experiments—one where two hundred and forty 12th grade EFL students were tested by the above form oral tests (oral interview, role play, reporting and group discussion) twice—once in laboratory conditions and again in their schools. Half of the students were tested in their schools by an integrated test (one tester, one event, multiple speech styles) and the other half by an isolated version of the test (four testers, four events, four speech styles). All the laboratory tests were done in the isolated manner.

Comparison of the two procedures with the laboratory test showed that there were differences in students' scores between the two versions. Specifically, it was found that students' scores within the integrated version were more similar to one another than their scores on the isolated version. The differences among the different speech styles seemed to wash off in the integrated version. It was important to determine what the source of the difference was: the tester who was influenced by the test taker's performance on the first subtest, or the test taker who behaved the same in all these subtests. The second experiment describes the results of a study where this question was investigated. Fifty 12th grade students were tested twice—once by four isolated tests and once by an integrated test. All the taped tests were then re-rated, by additional raters, and the battery test was divided into separate segments, each representing one speech style and then re-rated. The results of the analysis will be reported and recommendations as to the format and procedure of overall oral proficiency tests will be made.
F. Sang, B. Schmitz, H. J. Vollmer, J. Baumert, P. M. Roeder

Models of Second Language Competence: A Structural Equation Approach

New evidence against a general language proficiency factor (GLPF)

In the discussion about the structure of second/foreign language ability a certain consensus seems to have been reached during the past years: on the other hand, neither the unitary competence hypothesis (UCH) in its strong form as advocated by Oller and others (cf., for example, Oller 1976) nor a multidimensional model of L2-competence along the lines of the four communicative skills could be evidenced by the data presented in different parts of the world (cf. Vollmer, 1982, 1983). Recent developments in language testing research, however, clearly demonstrated that a plurality of factors underlie language proficiency (cf. Bachman/Palmer 1981, Upshur/Homburg 1983 and others) - independent of the question how those could be interpreted and labelled adequately. As a result it became clear that there is an urgent need for a more adequate theoretical basis to explain the accumulating body of data (cf. Cummins 1981, Oller 1983).

It could be shown that the results of the most important empirical studies up to 1980 were largely dependent on the method of statistical analysis applied, namely that of the classical factor analysis, used as a hypothesis-generating device (cf. Sang/Vollmer 1978, Vollmer/Sang 1983). Accordingly, it was agreed by some researchers that the confirmatory factor analysis (allowing for a statistical comparison between a predicted model and the results actually achieved) was a more powerful instrument and should be used in the future (cf. Bachman/Palmer 1983, Vollmer/Sang 1983).

With respect to this background the paper presents new empirical evidence against the assumption of a GLPF, in favor of a multiple factor model of second language ability. On the basis of confirmatory factoring techniques a three-dimensional model will be suggested, which seems to be theoretically meaningful and also best fits our data (n = 12.252). The three components are labelled as follows:

- basic knowledge,
- integration of basic knowledge elements,
- interactive, communicative use of language,

The relationship of these components is considered to be one of stepwise implication.

Moreover, the paper will address the problem, in how far differences in cognitive outfit, namely achievement in L1, influences the three-dimensional structure of foreign language competence found so far. It will be
When periodic writing assignments serve as a testing device — i.e., a source for assessing or grading student performance in the classroom, teacher feedback on such assignments is provided with the intention of promoting improvement on subsequent similar tasks. In other words, the feedback data are expected to have an impact on the students' writing. Otherwise, the teachers would not take the time to generate this feedback. While the teachers' motives are well-placed, this paper questions the utility of such feedback as commonly provided. Reference is made to a growing research literature which indicates that first- and second-language writers either pay less attention to feedback than their teachers think, or pay attention but with dubious results (Marzano and Arthur 1977; Semke 1984; Robb, Ross and Shortreed 1984; Cohen 1984; Hayes and Daiker 1984).

The problem is partly a result of the nature of the feedback in that it is not as meaningful to the student writers as the teachers believe. The problem is also partly created by the learners themselves in that they reject or only partially accept even the best of feedback. It appears that students have both affective and cognitive reasons for doing this, in accordance with their own personal characteristics.

The paper considers the teacher's role as tester/evaluator in providing meaningful feedback and in ensuring that the feedback is utilized effectively by the students.

Cohen, Andrew D. The processing of feedback on second-language compositions. Paper presented at the Colloquium on Research on Learner Strategies, 18th TESOL Convention, Houston, TX, March 6-11, 1984.


Robb, Thomas, Ross, Steven, and Shortreed, Jan. What's the best way to correct compositions? The Language Teacher, 1984, 8 (6), 7-8.

shown that, according to our data, high achievers in German as L1 (n = 1056) as much as low achievers (n = 1127) seem to be governed in their foreign language learning by those three factors identified, but with differing patterns of loadings.

Finally, the paper will demonstrate that different sets of instructional strategies in second language learning and teaching have a specific and differentiating influence on various aspects of L2 achievement. In other words, a traditional approach has a strong influence on the development of basic elements of knowledge in English as L2, whereas it has no visible effect on the integration of such knowledge elements and a negative effect on the interactional use of the language learned. A more modern, interactive approach, on the other hand, favors good achievement in all aspects measured, particularly in the communicative areas.

In spite of the limitations of the data-set analysed, these results will all be interpreted as strong and further evidence against a GLPF (and its educational implications) and for a multiple factor model of second ability as outlined above.
The Promise and Threat of Computerized Adaptive Assessment of Reading Comprehension

The general goal of this presentation is to examine the potential of computerized adaptive testing (CAT) with respect to two fundamental issues in language testing research: construct validity and test method effects. To this end the presentation will focus on certain major strengths and weaknesses of computerized adaptive assessment of reading comprehension, which is the language skill area most often addressed in research on CAT.

With respect to the issue of the construct validity of assessment of reading comprehension, it will be argued that the most important potential strength and weakness of CAT involve its heavy dependence on the model of measurement theory often referred to as item response theory (or more generally as latent structure analysis). In particular, perhaps the main strength of such a version of CAT is that it must be based upon a sound and explicit analysis of different ability levels in reading comprehension --i.e. upon a sound construct or theory of reading comprehension. Perhaps the main weakness of this version of CAT is that the construct to be measured must, according to item response theory, be unidimensional--i.e. largely involve only one major factor. Not only is it difficult to maintain that reading comprehension is a unidimensional construct (for example, to ignore the influence of world knowledge); it is also difficult to understand how CAT could serve useful diagnostic and achievement purposes if reading comprehension is assumed to be unidimensional and hence neither decomposable into meaningful subparts nor influenced by instruction.

With respect to the issue of test method effects in assessment of reading comprehension, it will be argued that CAT offers a unique strength and weakness. The main strength may be the potential for use of assessment techniques that are not easily administered in a non-CAT format, such as more process oriented, interactive and nonintrusive tasks that can be integrated into broad and thematically coherent language use/learning activities (for example, intelligent tutoring systems). The main weakness may be the tendency among both large-scale and small-scale test developers simply to mechanize existing product-oriented reading comprehension item types (for example, a thematically incoherent mix of short reading passages each followed by one or two multiple-choice questions) in an effort to enhance the efficiency of test administration and assure conformity to crucial assumptions of item response theory.
Computerized Rasch Analysis of Item Bias  
in Selected ESL Subtests

The purpose of this study is to utilize latent trait measurement  
in the form of the Rasch one-parameter logistic model (Rasch 1980) to  
identify item bias in ESL subtests of listening, reading, and grammar.

The Rasch model has been successfully utilized for such varied  
purposes as test item calibration (Englehard 1980), learner-specific item  
analyses (Rasch 1966), item banking (Rubin and Mott 1983), test linking,  
and computerized testing (Henning 1984). The Rasch model can also assess  
the precision with which a test measures individuals at a variety of  
ability levels, through the calculation of an "information curve"  
(Hambleton and Cook 1977).

Test bias is a recurring concern in the profession. Houts (1977),  
Oller and Perkins (1978), and Leach (1979), in evaluating contemporary  
educational tests, point up difficulties ranging from improper evaluation  
of nonnative speakers of English, to exacerbation of social inequities.  
Cultural and socio-economic bias in tests have been reported on by Nakono  
(1977), Garcia (1979); instances of cultural bias in language proficiency  
exams have been discussed by Mohan (1979) and Briere (1972). Moreover,  
test form (Farhady 1979) and test anxiety (Madsen 1982) have been seen as  
sources of ESL test bias.

The study reported in this paper utilizes sets of Rasch-generated  
test-item analyses to assess the bias in ESL battery subtests. Subjects,  
125 ESL students of beginning to intermediate English proficiency at a  
pre-university English language institute, were administered a five-part  
placement battery. Three multiple choice subtests—grammar, listening  
comprehension, and reading comprehension—are being Rasch analyzed  
utilizing Wright's Microscale Plus program for the microcomputer (1984).

The study is designed to disclose the relative bias of items. Two  
hypotheses will be tested: 1) there are identifiable patterns of items  
that generate a poor-fit statistic, which can be accounted for by student  
language background, 2) similarly, there are identifiable patterns of  
items that produce a poor fit statistic, regardless of subjects' language  
background.
The Oral Proficiency Interview has an established reputation as a classic adaptive test of speaking in a foreign language. The trained interviewer creates the test by assessing the speaker's response at each step before advancing to the next question. An American Council on the Teaching of Foreign Languages (ACTFL) project seeks to develop a computer adaptive reading and listening test that replicates the tenets of the Oral Interview for receptive skills. The adaptive potential of the computer will simulate the decisions made by the human interviewer. Based upon previous answers, the computer will move the student to a higher level or a lower one as responses warrant.

The first phase of the project involves the design of test items which are matched to passages typed according to the proficiency levels as delineated by ACTFL and the government agencies. These will be developed and piloted in a paper/pencil format before any programming is done, but the item types will be chosen which lend themselves best to a computer delivery system. The listening test will use an interactive video program so that students experience authentic listening materials.

Test developers are particularly interested in determining if item response theory will serve to identify items at the various ability levels of the students. The advantages of a computer adaptive test for receptive skills are numerous: efficiency (the reader/listener need not go through lengthy sequences too low or too high for level); security (each test will be potentially a different assortment of items); final rating (the program can account for the patterns of performance that comprise a reader/listener's "floor" and "ceiling" and render a final judgment).
Item Response From Pretests in Mixed Ability Groups

Since 1983 two different types of secondary education in The Netherlands have been brought together in one single system. However, because of the width of the ability range, examinations at two different levels are offered at the end of the curriculum which is identical for all. Partial overlap of the examinations at both levels introduced for efficiency reasons is justified by the overlap of the ability ranges at both levels and permits equating procedures. As the choice of examinees for either level of assessment is based on part of the examination procedure, items are pretested on samples of 300 to 1000 subjects from the total group but have to be evaluated as to their estimated appropriateness at one of either levels.

The usefulness of the Rasch model is demonstrated and the appropriateness of the tests for the target levels is evaluated.
Using Technology to Assess Second Language Proficiency Through Cloze-Elide Tests

Among important developments in the testing of second language proficiency is the wide interest in integrative tests of language (cloze and dictation tests) in contrast to more traditional discrete point tests (e.g., grammar and vocabulary). Standard cloze tests ordinarily involve altering a prose passage by randomly deleting words and then requiring the examinee to supply the missing words, relying upon the contextual meaning afforded by the remaining "mutilated" text. This method, although providing highly reliable and valid assessment is of limited use in large-scale testing programs because it is laborious and expensive to score. The author is exploring a new approach to second language assessment called cloze-elide testing. In this method extraneous words are randomly inserted in prose passages, and the examinee is required to strike out, or elide, these words thereby restoring the text to its original form. A specially designed answer sheet permits scoring these exercises by means of optical scanners, with the benefit of rendering this form of cloze-testing suitable for large scale testing programs. A program of research has been initiated to evaluate the validity of cloze elide tests for assessing language competency. Among the projects now under way are: (1) an evaluation of cloze-elide testing of ESL students supported by the Test of English as a Foreign Language (TOEFL program); (2) an investigation of the construct validity of cloze-elide testing for possible inclusion in NAEP (at age/grade 9/3, 13/7, and 17/11); (3) a study of the utility of cloze-elide exercises for a new lower level TOELC (Test of English in International Communication); and (4) a project to develop a microcomputer version of cloze-elide testing. A prototype exercise, administered through the IBM/PC by means of a light pen, has been recently developed and will be demonstrated.
Computer Text-Analysis in ESL Composition Teaching and Testing

Since 1981, when CSU acquired and began to adapt Bell Laboratory's Writer's Workbench (WWB) text-analysis software for collegiate use, various research projects have studied the efficiency of text-analysis in teaching composition and the relationship between text-analysis data and holistic scoring of student essays.

Following a brief introduction to the WWB, this paper will report on the following research:

Keifer and Smith

1984: a study that tested the effectiveness of the WWB text-analysis program on the editing and revision skills of composition students using the WWB.

Reid and Findlay

1984: a study that evaluated to what degree the stylistic measurements in the WWB programs correlate with student essay quality, as measured by holistic scores.

Reid

1985: a study funded by the Exxon Education Foundation that examined the impact of the WWB on international student writing skills.

Cumulatively, these research projects have resulted in the refinement of the CSU WWB standards for stylistic quality in different kinds of student writing, and in a wealth of seminal information concerning the ways in which computer text-analysis can measure essay quality.
The Correlation Between Machine and Human Ratings
of the Pronunciation of TSE Reading Passages

The purpose of this research is twofold. First, it is to demonstrate the current applicability of linear predictive coding to the characterization and evaluation of pronunciation. Second, it is to justify suggestions to further expand speech processing research into testing and evaluation.

The data for this research consisted of tapes provided by ETS of 20 nonnative English speakers. Each examinee read aloud a 125 word passage that appeared on a disclosed form of the TSE. The samples were chosen by ETS to provide a broad range of levels. The tapes were processed through an analog to digital converter, taking 10,000 samples per second for a frequency range of 0-5000 Hz. A Harris 800 super minicomputer was used for the analysis.

Linear Predictive Coding (LPC) was designed as a method of data compression. It provides a model of the vocal tract, dividing it into 10 sections. By combining the coefficients of the areas of each section with pitch period data, we are able to uniquely define each phoneme in a passage. By using standard statistical packages, we are able to automatically match patterns with a passage, and compare these patterns to those of other passages. This provides a quantifiable means for characterizing and evaluating pronunciation. The resulting graphs show the closeness of fit between the patterns recognized by the machine and the perceptions expressed by professional raters. Additional filters could enhance the applicability of LPC for this purpose. These will also be discussed.
The Use of Computers to Identify Two Different Stages of Language Acquisition for Thai Students

In recent years, construct validity has been the highlight of language testing research. Researchers in the field now seem to agree that communication is the goal of language teaching. According to the research studies conducted by Bachman and Palmer (1980), Upshur (1980) and Carroll (1980), there exist a general factor and additional factors. What remains for language testing researchers is to characterize the nature of communicative competence.

In second language acquisition it is believed that the learner tends to acquire the target language faster if one is exposed to the input that is appropriate to his acquisition stage (the $i+1$ level in Krashen's term). However, to find the acquisition stage to which a student belongs has not been possible, i.e., no testable or measurable way of accurately placing a student into the correct stage has been discovered. This study, therefore, aims at determining the stages of language acquisition to which Thai students belong.

This research aimed to place students into one of two broad acquisition stages: stage 1 (below level three) and stage 2 (between level three and level six). The subjects were 2801 secondary school students randomly selected from 17 schools in Bangkok, Thailand. Two pragmatic tests, one in the visual mode and the other in the auditory mode, were used. Each pragmatic test consisted of two subtests: one aimed at the first stage of language acquisition, the other at the second stage. One-way ANOVA was performed to help explain the relationship between English proficiency and the two stages of language acquisition. Discriminant analysis was used to distinguish the learners at the first stage of language acquisition from those at the second stage. The results seem promising and suggest that computers can help language testing researchers identify the stages of language acquisition of the learners.
This paper reports the findings of a factor analytic (FA) study undertaken to explore the dimensions which underlie second language proficiency. In contrast to previous FA studies which have reported that many types of ESL proficiency tests appear to measure a unidimensional attribute, this study's factor analysis revealed that two factors identified as an academic skills factor and a functional use factor accounted for all the common factor variance in 16 different ESL measures. In view of Chronbach's warnings about the "treacherousness" of factor analysis as a research tool, the identification of the factors as academic skills and functional use was treated as tentative and subject to further investigation. Therefore, to verify the labels "academic" and "functional", factor scores were computed for each factor. These scores, composite scales created by multiplying a subject's standardized score (a "z" score) on each measure by the factor score coefficient (a weighting) and then summing the weighting-times-z-score products, are indices of the theoretical dimensions associated with the factors. In this case, they were assumed to be indices of the academic skills and functional use aspects of measures used in the study. These indices were then correlated with demographic variables, years of residence in an English speaking country and number of college credits completed. It was hypothesized that if the two factors had been appropriately identified the relationships between each factor and each demographic variable would vary in a specific way. The relative strength and direction of the pearson product moment correlation coefficients provided evidence to support the identification of the factors as an academic skills factor and a functional use factor. The relevance of these findings to the acceptance of ESL tests as overall measures of general proficiency will be discussed.