DOCUMENT RESUME

ED 223 103 FL 013 318

AUTHOR Palmer, Adrian S., Ed.; And Others
TITLE The Construct Validation of Tests of Communicative Competence.
INSTITUTION Teachers of English to Speakers of Other Languages.
PUB DATE 81
NOTE 171p.; Includes proceedings of a colloquium at TESOL (Boston, MA, February 27-28, 1979). For individual papers, see FL 013 319-329.
AVAILABLE FROM TESOL, 202 D.C. Transit Building, Georgetown University, Washington, DC 20057.
PUB TYPE Collected Works - Conference Proceedings (021) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Communicative Competence (Languages); English (Second Language); Higher Education; Language Proficiency; Language Research; *Language Tests; Reading Tests; *Second Language Learning; Speech Communication; Speech Tests; Testing; *Test Validity

ABSTRACT

This collection, including the proceedings of a colloquium at TESOL 1979, includes the following papers: (1) "Classification of Oral Proficiency Tests," by H. Madsen and R. Jones; (2) "A Theoretical Framework for Communicative Competence," by M. Canale and M. Swain; (3) "Beyond Faith and Face Validity: The Multitrait-Multimethod Matrix and the Convergent and Discriminant Validity of Oral Proficiency Tests," by D. Stevenson; (4) "Convergent and Discriminant Validation of Integrated and Unitary Language Skills: The Need for a Research Model," by R. Clifford; (5) "Structure of the Oral Interview and Content Validity," by P. Lowe, Jr.; (6) "A Study of the Reliability and Validity of the Ilyin Oral Interview," by A. Engelskirchen, E. Cottrell, and J. Oller, Jr.; (7) "Inter-rater and Intra-rater Reliability of the Oral Interview and Concurrent Validity with Cloze Procedure," by E. Shohamy; (8) "Assessing the Oral Proficiency of Prospective Foreign Teaching Assistants: Instrument Development," by F. Hinofotis, K. Bailey, and S. Stern; (9) "Measurements of Reliability and Validity of Two Picture-Description Tests of Oral Communication," by A. Palmer; (10) "An Experiment in a Picture-Stimuli Procedure for Testing Oral Communication," by L. Bachman; and (11) "A Multitrait-Multimethod Investigation into the Construct Validity of Six Tests of Speaking and Reading," by L. Bachman and A. Palmer. (AMH)

The Construct Validation of Tests of Communicative Competence

Including proceedings of a colloquium at TESOL '79, Boston February 27-28, 1979

Edited by

Adrian S. Palmer
Peter J. M. Groot
George A. Trosper

Teachers of English to Speakers of Other Languages
Washington, DC, USA
1981

DOCUMENT RESUME

ED 223 103 FL 013 318

AUTHOR Palmer, Adrian S., Ed.; And Others
TITLE The Construct Validation of Tests of Communicative Competence.
INSTITUTION Teachers of English to Speakers of Other Languages.
PUB DATE 81
NOTE 171p.; Includes proceedings of a colloquium at TESOL (Boston, MA, February 27-28, 1979). For individual papers, see FL 013 319-329.
AVAILABLE FROM TESOL, 202 D.C. Transit Building, Georgetown University, Washington, DC 20057.
PUB TYPE Collected Works - Conference Proceedings (021) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Communicative Competence (Languages); English (Second Language); Higher Education; Language Proficiency; Language Research; *Language Tests; Reading Tests; *Second Language Learning; Speech Communication; Speech Tests; Testing; *Test Validity

ABSTRACT

This collection, including the proceedings of a colloquium at TESOL 1979, includes the following papers: (1) "Classification of Oral Proficiency Tests," by H. Madsen and R. Jones; (2) "A Theoretical Framework for Communicative Competence," by M. Canale and M. Swain; (3) "Beyond Faith and Face Validity: The Multitrait-Multimethod Matrix and the Convergent and Discriminant Validity of Oral Proficiency Tests," by D. Stevenson; (4) "Convergent and Discriminant Validation of Integrated and Unitary Language Skills: The Need for a Research Model," by R. Clifford; (5) "Structure of the Oral Interview and Content Validity," by P. Lowe, Jr.; (6) "A Study of the Reliability and Validity of the Ilyin Oral Interview," by A. Engelskirchen, E. Cottrell, and J. Oller, Jr.; (7) "Inter-rater and Intra-rater Reliability of the Oral Interview and Concurrent Validity with Cloze Procedure," by E. Shohamy; (8) "Assessing the Oral Proficiency of Prospective Foreign Teaching Assistants: Instrument Development," by F. Hinofotis, K. Bailey, and S. Stern; (9) "Measurements of Reliability and Validity of Two Picture-Description Tests of Oral Communication," by A. Palmer; (10) "An Experiment in a Picture-Stimuli Procedure for Testing Oral Communication," by L. Bachman; and (11) "A Multitrait-Multimethod Investigation into the Construct Validity of Six Tests of Speaking and Reading," by L. Bachman and A. Palmer. (AMH)

Library of Congress Catalog Card No. 81-53026

Copies available from:

TESOL
202 D.C. Transit Building
Georgetown University
Washington, DC 20057

Table of Contents

Preface

This collection of papers is directed essentially to the language testing professional and others with theoretical and research interests. Such readers will find here a fairly thorough consideration of one previously neglected area of language testing theory: the construct validation of tests of communicative competence. However, readers without a strong background or interest in research will find that they have not been neglected. The introductory paper, by Palmer and Groot, written specifically for this volume, provides the necessary orientation to understand the nature of the problems addressed and the conclusions reached. Moreover, many of the papers contain descriptions of test administration and discussions of tests' strengths and weaknesses; these should offer increased insight for the classroom test administrator into factors affecting the choice, use, and interpretation of tests.

As the subtitle indicates, this volume contains the proceedings of a colloquium. However, the contents are not limited to the papers given at Boston in 1979. This volume sets out, in fact, to trace one cycle of research in language testing: the original voicing of concern over the lack of adequate construct validation of any oral proficiency test in use; the consultation among concerned researchers leading to the Boston colloquium at which the necessary new research was outlined; and the report of that new research as actually conducted and the conclusions reached — with a glimpse of directions for a new cycle of research.

Foreword

In August of 1978, at the Fifth International Congress of Applied Linguistics in Montreal, Peter J. M. Groot voiced a concern that while the need for oral proficiency testing (and therefore general attention to it) was increasing, very little attention had been paid to the question of construct validity. He suggested that the 1979 TESOL convention in Boston would be a good opportunity for researchers interested in the validation of oral tests to meet and discuss this issue. It seemed probable that such contact would stimulate the necessary empirical research. Groot and Adrian Palmer began contacting researchers in the field and found that there was indeed considerable interest in this subject. With the support of the TESOL organization (Teachers of English to Speakers of Other Languages), they arranged to hold a colloquium during the first days of the convention.

At the 1979 Boston colloquium, more than a dozen papers were presented and discussed over a two-day period, and several hours of general planning sessions were held.[1] The results of this colloquium have been valuable in both general and very concrete ways. In general, 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 the others. In addition, the colloquium produced three more concrete outcomes.

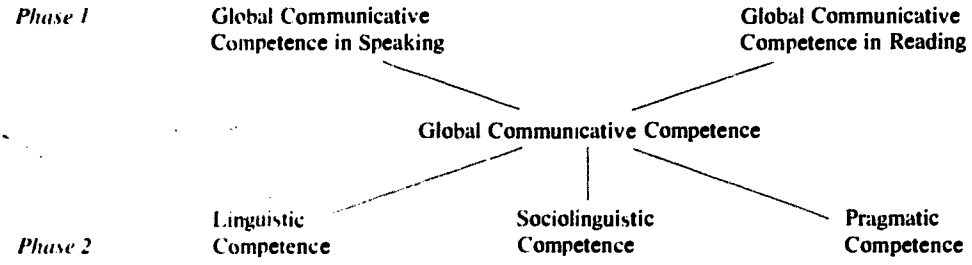
The first was the outlining of a long-range empirical investigation into the construct validity of tests of communicative competence. This investigation was to proceed in two phases. Phase 1 was to define two maximally distinct global areas of language use and to seek evidence for the construct validity of tests of these areas. If the Phase-1 study provided evidence of such validity, the second phase of the investigation would be undertaken. Phase 2 would be an investigation of the construct validity of tests of the *components* of communicative competence. Anticipating this phase, colloquium participants developed provisional definitions of these components. The two phases in the investigation are pre-

[1] Colloquium participants, including authors and attendees, were Lyle F. Bachman, Kathleen M. Bailey, Michael Canale, Brendan Carroll, John L. D. Clark, Ray T. Clifford, Elinore Cottrell, Alan Davies, Alice Engelskirchen, Peter J. M. Groot, Deborah Hendricks-Sanchez, Frances B. Hinofotis, Donna Ilyin, Marianne Johnson, Randall L. Jones, Dale Lange, Pardee Lowe, Jr., Harold S. Madsen, John W. Oller, Jr., Adrian S. Palmer, Meredith Pike, Stephen B. Ross, George Scholz, Elana Shohamy, Randon Spurling, Charles Stansfield, Susan L. Stern, Douglas K. Stevenson, Merrill Swain, and Lela Vandenburg.

sented graphically in Figure 1 and the provisional definitions are given in an appendix to this foreword. (It was decided soon after the colloquium to drop the provisionally-defined fluency component, since it was incompatible with important testing methods, e.g., discrete-point and multiple choice. This component therefore does not appear in Figure 1.)

FIGURE 1

A Plan for a Two-Phase Investigation into the Construct Validity of Tests of Communicative Competence.



The second concrete outcome was the development of a specific design for the Phase-1 study. Decisions made included adopting global definitions of communicative competence in speaking and reading, determining which types of tests should be included, selecting appropriate tests where they already existed, and deciding on specifications for those tests which had to be developed. Two of the participants in the colloquium, Lyle Bachman and Adrian Palmer, agreed to carry out the Phase-1 study and to present the results at a second colloquium during the 1980 TESOL convention in San Francisco. The last paper in this volume reports the results of this study.

The third concrete outcome of the Boston colloquium is the publication of the present volume.

APPENDIX

Provisional definition of communicative competence in speaking

1. Ability to produce spoken language exhibiting control of the linguistic rules employed by the speakers of a given dialect or set of dialects. Control consists of breadth (range of structures attempted) and accuracy (degree to which the structures are produced correctly). Areas of linguistic control are phonology, morphology, and syntax.
2. Ability to produce spoken language exhibiting control of the socio-linguistic rules employed by the speakers of a given dialect or set of dialects. Sociolinguistic rules consist of conventions for producing textually cohesive speech, speech in an appropriate register, and speech incorporating appropriate cultural references. Control consists of breadth (range of

language-use situations in which the speaker is sensitive to prevailing standards in the above named areas) and accuracy (degree to which the language produced conforms to prevailing standards).

3. Ability to produce spoken language exhibiting control of the pragmatic rules employed by the speakers of a given dialect or set of dialects for communicating the types of messages required of these speakers. Pragmatic rules are conventions relating the form of an utterance to the intended meaning. Important factors in pragmatic competence are extent of vocabulary, and accuracy of pronunciation. Control consists of breadth (range and complexity of messages communicated) and accuracy (degree to which the language produced communicates correctly the details of the content).
4. Ability to produce spoken language fluently. Fluency consists of overall quantity of production and tempo of production. Control of overall quantity of production consists of the ability to produce an amount of language within a limited period of time consistent with native speaker norms for the type of message communicated. Control of tempo of production consists of the ability to maintain, confidently, a pace of rhythm consistent with norms for native speakers of a given dialect or set of dialects.

Provisional definition of communicative competence in reading

1. Ability to react to the linguistic rules manifested in the written language. Ability to react consists of breadth (range of structures reacted to) and accuracy (degree to which the reactions conform to prevailing standards). Areas of linguistic control are morphology and syntax.
2. Ability to react to the sociolinguistic rules employed in given written dialects or sets of dialects. Sociolinguistic rules consist of conventions used in the production of cohesive text, conventions used in the production of text in a register appropriate to the particular aims and modes of written discourse, and conventions for the incorporation of appropriate cultural references. Control consists of breadth (range of aims and modes for which the reader is sensitive to prevailing standards) and accuracy (degree to which the reactions conform to prevailing standards).
3. Ability to react to the pragmatic rules employed in a given written dialect or set of dialects. Pragmatic rules are conventions relating the form of a text to the intended meaning. Important factors in pragmatic competence are extent of passive vocabulary and knowledge of conventions relating linguistic units to their orthographic forms. Control consists of breadth (range of messages reacted to) and accuracy (degree to which the reactions conform to prevailing standards).
4. Ability to react to the written language fluently. Fluency consists of quickness of response to written material (degree to which speed of response conforms to prevailing standards).

*An Introduction**

Adrian S. Palmer
University of Utah
Peter J. M. Groot
University of Utrecht

The process of test validation is complex, and the papers in this volume address a particular problem — the validation of tests of communicative competence — from a variety of perspectives and in varying degrees of technicality. To those already familiar with the literature on validity, the papers need no introduction. However, to readers trying to gain familiarity with the concept, the number of new ideas introduced and technical terms used might prove frustrating. The first part of this paper, therefore, provides an introduction to validity for such readers, and the second offers a brief synopsis of the remaining papers in this volume.

Introduction to Test Validity

Validity: the concept

Validity is a frequently misunderstood concept. It is often erroneously believed that a test is simply valid or not valid, as if validity were a property of the test itself. In fact, as Cronbach has pointed out, one does not validate a test. One validates "an interpretation of data arising from a specified procedure" (Cronbach, 1971: 477). The elements affecting validity include, among others, the test itself, the setting in which the test is administered, the characteristics of the examiner, and the inferences intended to be drawn from the test. However, it should be noted that, in the literature, the word 'test' is frequently used to refer to the combination of the test itself (including setting, examiner, etc.) and the inferences drawn from scores on it. "The validity of a test" then *can* have meaning — as long as the distinction between the two uses is kept clearly in mind. Still, "it is incorrect to use the unqualified phrase 'the validity of the test.'

*We would like to thank George A. Trosper for his comments on this paper.

No test is valid for all purposes, for all situations, or for all groups of individuals." (APA, 1974, 31)

The general purpose of the validation procedure is, then, to investigate the extent to which inferences can properly be drawn from performance. The process of collecting evidence of the extent to which such inferences are warranted is called validation.

Kinds of validation

Of the several ways of evaluating validity, the three most important are discussed below: content validation, criterion-referenced validation, and construct validation.

Content validation. Content validation is the process of investigating whether the selection of tasks one observes in a test-taking situation is representative of the larger set (universe) of tasks of which the test is assumed to be a sample. For example, if a test is designed to measure "ability to converse in a foreign language" yet requires the testee only to answer yes-no questions, one might doubt that this single task is representative of the sorts of tasks required in general conversation, which entails operations like greeting, leave-taking, questioning, explaining, describing, etc. The paper by Lowe and that by Stevenson in this volume address the issue of content validity in detail, and the one by Palmer addresses it peripherally. Therefore, it will not be dealt with further in this introduction.

Criterion-referenced validation. Criterion-referenced validation is the process by which one "compares test scores, or predictions made from them, with an external variable (criterion) considered to provide a direct measure of the characteristic of behavior in question" (Cronbach, 1971: 444). The "criterion" in a criterion-referenced validation of a test is frequently simply another test or tests; but grade point averages, and other types of numbers not derived from anything generally considered to be a test, are also often available.

A study of criterion-referenced validity may be undertaken for the purpose of establishing either *predictive* validity or *concurrent* validity. A test has predictive validity when it can be used to make a prediction about a future event or state, e.g., success or perseverance in a course of study. Concurrent validity refers to the substitutability of a new test for one already in use, in order to save time and costs in administration and/or scoring.

It is important to note that in criterion-referenced validation knowing exactly what a test measures is not crucial, so long as whatever is measured is a good predictor of the criterion behavior. For example, a score on a translation test from a foreign language to English might be a very good predictor of how well a student would perform in courses in an English-medium university — even though it might not be at all clear exactly what the translation test measured: his knowledge of English, sensitivity to the foreign language, ability to

translate, perseverance, or some combination of these or other abilities. The test could have criterion-referenced validity whether or not these abilities had causal relevance to the student's passing courses in an English-medium university.

It is, in fact, possible for scores on tests of two distinct abilities to correlate highly without any actual causal relation between them. For example, let us assume that the scores on tests of reading ability and physical strength for young children are highly correlated: the higher the score of a child on one test, the higher the score that can be expected for him on the other test. Were one to use one test (say, of reading) as the criterion for evaluating the other test (of physical strength), one could conceivably claim high concurrent validity; but this could certainly *not* be used as evidence that the strength test actually measured reading ability. (What such a study would perhaps indicate is that for young children in school, both reading ability and physical strength are functions of underlying variables, such as age.) A similarly high correlation might also occur simply because two abilities had been taught to an experimental population in a single curriculum.

Putting aside for the moment our uncertainty as to what a test is measuring in any given case, let us turn our attention to the criterion. Many authors (e.g., Anastasi, 1950; Ebel, 1965) have pointed out the possible absence of valid criterion measures. This is demonstrated clearly in language testing, where "the question of what it is to know a language is not yet well understood and consequently the language proficiency tests now available and universally used are inadequate because they attempt to measure something that has not been well defined" (Jacobovits, 1970; cf. also Oller, 1973; Groot, 1975; Peterson and Cartier, 1975).

For example, as Upshur (1976) has pointed out, grammar items of the sort used in standardized tests such as the Test of English as a Foreign Language and the Michigan Test draw not only upon the student's knowledge of grammar (however that might be defined) but also on lexical knowledge and knowledge of the world in general. To use a test composed of such items as the criterion in a validation study is to place one's faith in a test which may not itself be a valid measure of the construct "knowledge of grammar," even if the test is standardized and widely respected. Ebel (1965) agrees: "The difficulties and uncertainties in getting directly valid *criterion* measurements are exactly as serious as those of obtaining directly valid test scores. In fact, the two problems are almost identical." As a consequence, criterion-referenced validation in the strictest sense of the term may not be possible because we in language testing — like professionals in other areas of education — do not always have one or more external variables (criterion measures) which we can demonstrate to be valid measurements of the psychological property (ability or trait) we are interested in.

The best one may be able to hope for in criterion-referenced validation is "successive approximation" to criterion validity. By this is meant that, in a validation study, the chances of not having measured the attribute one is after

with one's test become smaller and smaller each time one obtains a high correlation with another test designed to measure the same attribute. However, Cronbach and Meehl (1955) suggest that this process leads to "infinite frustration," pointing out that even if there were a valid criterion, a high correlation between a test and a criterion would not — as demonstrated by the examples at the beginning of this section — tell us much about what the scores on the test *mean*. This last objective is the goal of construct validation.

Construct validation. Construct validation is a process of investigating what a test measures. In education, this is usually one or more psychological properties (including what we have been calling "abilities").[1] For example, if it is claimed that a test measures "knowledge of grammar," one should be able to demonstrate that one can measure knowledge of grammar (as a psychological property) to a certain extent independently of other purported psychological properties such as "knowledge of vocabulary," "knowledge of the writing system," "ability to reason verbally," etc.

In construct validation, one validates a test not against a criterion or another test, but against a theory. To investigate construct validity, one develops or adopts a theory which one uses as a provisional explanation of test scores until, during the procedure, the theory is either supported or falsified by the results of testing the hypotheses derived from it. This sequence, common to all empirical research, will often be cyclical because, as Fiske (1971: 272) explains: ". . . concepts guide empirical research and empirical findings alter concepts. This interaction is the essence of science."

The construct validation of communicative competence

There are a number of different procedures for investigating construct validity (Cronbach, 1971). Two of these are described here because of their relevance to the research studies proposed at the colloquium in Boston. The first, a fairly general procedure, follows quite directly from the brief discussion of construct validation above. The second, a more specific procedure called multitrait-multimethod convergent-divergent validation, is employed in several of the papers in this volume.

A general procedure. One general procedure for investigating construct validity consists of five steps: defining what traits one is trying to measure, operationalizing the definitions by means of tests, stating hypotheses about the relationships between subjects' scores on the various tests, administering and scoring the tests, and comparing the obtained results with the hypothesized results. In this section, we illustrate these steps as applied to a hypothetical study of communicative competence in speaking.

[1] Various terms have been used for such properties, including "construct," "psychological property," "mental ability," and "trait." While distinctions might be made among these terms, they are used more or less interchangeably in this paper.

Canale and Swain (in this volume) have postulated three sets of factors contributing to communicative competence: linguistic factors (control of grammar, lexicon, and phonology), sociolinguistic factors (control of socio-cultural rules and discourse rules), and certain strategic factors (such as flexibility in choosing between alternative approaches to communication). As the first step in our construct validation procedure, these factors could be adopted as components constituting a provisional definition of communicative competence.

The second step in this procedure is to locate existing tests or develop new ones to operationalize the provisional definition. In our hypothetical study, an existing test, the Foreign Service Institute (FSI) oral interview (FSI, 1979), might serve to operationalize the general construct "communicative competence in speaking." One might also develop or locate tests which operationalize each of the individual components in the general "communicative competence in speaking" construct (i.e., linguistic competence, sociolinguistic competence, and strategic competence).

The third step is to form hypotheses and make predictions. In the study we are considering, predictions grounded in theory would be made about the magnitude of the correlations between subjects' scores on the FSI oral interview and their scores on tests of individual components in the model, such as those in the following list.

correlations between FSI interview scores and scores on tests of linguistic competence	$> .70$	
correlations between FSI interview scores and scores on tests of sociolinguistic competence	$> .50$	$< .70$
correlations between FSI interview scores and scores on tests of strategic competence	$> .30$	$< .50$
correlations between FSI interview scores and scores on a test of a presumably independent (unrelated) competence, such as mathematical ability	$> .20$	

The fourth step in the procedure is to adminster the tests to a selected experimental population.

The fifth and last step is to compare the obtained results with those which ought to be obtained (assuming the model is accurate) if the tests measure what they are supposed to measure. In the present case this requires calculating the correlations listed in the third step above and comparing the values actually obtained against those hypothesized. Failure of the obtained correlations to conform to the predicted pattern would lead to the development of a new model

(theory), or of tests which might be better operationalizations of the construct as previously defined, or of both.

The multitrait-multimethod convergent-divergent construct validation procedure. There is a specialized construct validation procedure called the multitrait-multimethod convergent-divergent procedure. (The meanings of "multitrait," "multimethod," "convergent validity," and "divergent validity" will be discussed in subsequent paragraphs of this section). It is central to some of the research described in this volume and much of that projected for the future. The procedure was first described by Campbell and Fiske (1959) and was first recommended for use in the evaluation of language proficiency measures by Stevenson (1974). It is based on the assumption that a test score is a function both of the trait the test measures and of the method by which it is measured. For example, on a multiple-choice test of grammar, subjects' scores would be due in part to their ability to do multiple-choice tests (a component of method — something which one would *not* want to consider part of the psychological property "knowledge of grammar"). Two testees with equal knowledge of grammar but unequal testwiseness (knowledge of effective strategies for taking multiple-choice tests) would obtain different scores on the test.

In order to measure the relative contributions of trait (grammatical competence) and method (multiple-choice testwiseness), it is necessary, for statistical reasons, that two or more traits each be measured by two or more distinct methods.[2] It is for this reason that the procedure is called a multitrait-multimethod procedure. For example, one might measure each of the two traits "competence in grammar" and "competence in vocabulary" by means of two methods, a multiple-choice method and a fill-in-the-blank method, and then look for two types of validity.

The first type is *convergent validity*. The idea behind convergent validity is that persons scoring high on one valid test of a trait should also score high on a different valid test of the same trait. In the context of the example study described above, evidence of convergent validity would be a high correlation between scores on the two tests of the "grammar" trait (i.e., the grammar test using the multiple-choice testing method and the grammar test using the fill-in-the-blank method). Likewise, one would hope for a high correlation between scores on the two tests of the "vocabulary" trait. Low correlations would be grounds for questioning the convergent validity of these tests.

The second, far less frequently investigated, type of validity is *discriminant validity* — also called *divergent validity*. Simply put, a discriminant validation study looks for evidence that one trait can be measured separately from another. Again in the context of our example study, if "grammar" and "vocabulary" are independent traits one would not expect persons scoring high on grammar tests necessarily to score high on vocabulary tests also. If, in validation studies,

[2] The number of traits and methods required to produce *optimally* interpretable results is discussed by Alwin (1974) and by Althauser (1974).

scores on grammar and vocabulary tests were always highly correlated — no matter what single method was used to test both — this would be grounds for questioning either the discriminant validity of the tests examined in the studies or the distinctness of the traits labeled "grammar" and "vocabulary."

The *multimethod* component of the multitrait-multimethod procedure makes it possible to investigate convergent validity. Any difference between the tests of a trait may be attributable to the difference in the methods employed in the tests.

The *multitrait* component of the multitrait-multimethod procedure makes possible the investigation of discriminant validity, which requires measures of traits purported to be different — grammar and vocabulary, in the study described above.

Traditionally, designs for multitrait-multimethod convergent-divergent validation studies are displayed in a matrix. On one axis, the experimenter names the traits he is attempting to measure. On the other axis, he names the methods he will use to measure the traits. In each of the cells in the matrix, he names a particular test — which will be a combination of one trait and one method. Such a matrix for our example study is illustrated in Figure 1.

FIGURE 1

Multitrait-multimethod matrix for a hypothetical construct validation study

Traits \ Methods	Multiple-choice	Fill-in-the-blank
Grammar	Test #1: Multiple-choice test of grammar	Test #2: Fill-in-the blank test of grammar
Vocabulary	Test #3: Multiple-choice test of vocabulary	Test #4: Fill-in-the-blank test of vocabulary

In this particular study, evidence of convergent validity for the grammar tests would be high correlations between scores on tests #1 and #2. Evidence of convergent validity of the vocabulary tests would be high correlations between scores on tests #3 and #4. Evidence of discriminant validity would be low correlations between scores on tests #1 and #3 and tests #2 and #4 — and, of course, tests #1 and #4 and tests #2 and #3, which pairs share neither method nor trait.

Failure to find convergent validity. There are two likely reasons for failure to find convergent validity. One is that the methods used in the tests have

exerted so much influence on the test scores that they have obscured the effect of the trait one was trying to measure. For example, the effect of differences between multiple-choice testwiseness and fill-in-the-blank testwiseness might be the major influence on the scores.

The other likely reason is that the tests used to measure the trait were poorly constructed. If this seems probable, one could attempt to develop better tests and repeat the study.

Failure to find discriminant validity. If one fails to find evidence of discriminant validity, this also may be due to either of the reasons given in the previous section. Thus, in the hypothetical study we are examining, if the influence of the test methods is excessive, an actual difference between the "grammar" and "vocabulary" traits might be obscured. And, of course, poorly constructed tests will produce poor data in any study.

But there are additional possibilities. For instance, the traits one is trying to measure may not be "pure"; that is, each trait may actually consist of a number of subtraits (or components), some of which are common to both the hypothesized main traits. The effect on the test scores of these common subtraits might then be sufficiently strong to obscure the effects of whatever subtraits are unique to each main trait. For example, suppose one were to try to validate tests of competence in reading and writing yet failed to obtain evidence of discriminant validity. An explanatory hypothesis might be that both "reading" and "writing" traits share a number of subtraits in common, such as "grammar" and "vocabulary."

Yet another possibility is that the hypothesized traits are simply not independently measurable, at least not to the extent that evidence can be provided of discriminant validity. In this case, the experimenter must either rely on faith to justify his trait model or he must discard or revise it.

The Papers in This Volume

The papers in the volume fall into two general groups. Section I includes five papers on general approaches to oral proficiency testing: on the nature of communicative competence, on the philosophy of validation and its implications for the design of validation studies, on the implications of three validation studies viewed through the multitrait-multimethod convergent-divergent perspective, and on the content validity of the oral interview procedure. Section II includes six papers reporting on specific research into the reliability, validity, practicality, and use of oral tests.

In the first paper in Section I, Madsen and Jones present a profile for describing over a hundred oral proficiency tests they have collected. They discuss each of the categories in their classification and generalize about the relative amount of attention given to each category in oral testing as a whole.

Canale and Swain's paper presents a condensed version of their extensive

overview (Canale and Swain, 1979) of attempts in the literature to define "communicative competence" and suggests their own three-factor framework.

In his paper, Stevenson surveys various attitudes toward validation studies, criticizes many, and argues rather passionately for an attitude he calls "the spirit of validation."

In the fourth survey paper, Clifford examines examples of validation studies in the literature from the multitrait-multimethod convergent-divergent perspective and concludes that they consistently fail to provide evidence of construct validity for the traits the tests purport to measure. He attributes this to a failure to take into account (or control for) the effects of test method on test scores. Finally, he makes a number of specific recommendations for the design and implementation of validation studies.

In the final paper of the section, Lowe discusses an important type of validity which has been only defined in this introductory paper: content validity. He considers it in the context of the FSI oral interview — currently one of the most widely used and well-defined methods for assessing communicative competence in speaking.

The papers in Section II, as mentioned above, describe specific research. They deal not only with validity but also with important related issues: reliability (a prerequisite to validity), practicality of specific methods, procedures for developing and improving tests, alternatives to "direct" tests of communicative competence in speaking, and various uses of oral tests.

Engelskirchen, Cottrell, and Oller investigate the reliability and validity of one of the few widely publicized, readily available alternatives to the FSI oral interview: The Ilyin Oral Interview. Of particular interest in their study is the analysis of individual test items which take into account appropriateness or naturalness, a consideration often neglected by researchers bedazzled by numerical analysis.

Shohamy examines an FSI-type interview test for evidence of inter-rater reliability (the amount of agreement between different raters) and intra-rater reliability (the amount of agreement between a single rater's ratings on one occasion and his ratings on another occasion or under different circumstances) — two types of reliability which are important in any test scored by raters. She describes the basic procedure for training raters which led to her obtaining remarkably high reliability coefficients, and presents evidence of concurrent validity (agreement with cloze test scores).

The paper by Hinofotis, Bailey, and Stern is of both practical and theoretical interest. Of practical interest is the specific application for which they developed their test: screening foreign applicants for teaching assistantships on the basis of their oral proficiency in English. Of theoretical interest is the procedure used to develop the test and the analysis of the factors which contribute to the overall assessment of competence for this specific task.

The fairly bright picture of the state of oral testing painted so far is darkened

somewhat in the paper by Palmer. Describing the fall from favor of a once apparently promising procedure for measuring oral proficiency (a picture-description test), the paper details certain statistical improprieties in reliability studies and suggests test inadequacies related to content validity which limit the usefulness of this type of test.

On a more positive note, Bachman's paper, like that of Hinofotis, Bailey, and Stern, illustrates a practical application of oral testing: program evaluation. He discusses considerations in tailoring the content and method of oral tests to the evaluation of particular programs of instruction.

The final paper, by Bachman and Palmer, describes the results of the Phase-1 study as planned by the participants in the colloquium and carried out during the year following. Bachman and Palmer conclude that there is sufficient evidence of the existence of two distinct traits (communicative competence in speaking and in reading) to warrant further investigation into the possible components of communicative competence — Phase 2 of the planned investigation.

REFERENCES

Althauser, Robert P. 1974. Inferring validity from the multitrait-multimethod matrix: another assessment. In H. L. Costner, ed. *Sociological methodology 1973-1974*. San Francisco: Jossey-Bass.

Alwin, Duane F. 1974. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner, ed. *Sociological methodology 1973-1974*. San Francisco: Jossey-Bass.

American Psychological Association (APA). 1974. *Standards for educational and psychological tests*. Washington, D.C.: APA.

Anastasi, A. 1950. The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement 10*.

Campbell, D. T. and D. W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin 56*, 2.

Canale, M. and M. Swain. 1979. *Theoretical bases of communicative approaches to second language teaching and testing*. Toronto, Canada: The Ontario Institute for Studies in Education. (Mimeo)

Cronbach, L. J. 1971. Test validation. In R. L. Thorndike, ed. *Educational Measurement*, 2nd ed. Washington, D.C.: American Council on Education.

Cronbach, L. J. and P. E. Meehl. 1955. Construct validity in psychological testing. *Psychological Bulletin 4*.

Ebel, R. L. 1965. *Measuring educational achievement*. Englewood Cliffs, N.J.: Prentice Hall.

Fiske, D. W. 1971. *Measuring the concepts of personality*. Chicago, Ill.: Aldine Publishing Co.

Foreign Service Institute (FSI). 1979. *Testing kit: French and Spanish*. Washington, D.C.: Department of State.

Groot, P. J. M. 1975. Validation of language tests. In L. Palmer and B. Spolsky, eds. *Papers on language testing: 1967-1974*. Washington, D.C.: TESOL.

Jacobovits, L. A. 1970. *Foreign language learning: a psycholinguistic analysis of the issues*. Rowley, Mass.: Newbury House.

Oller, J. W., Jr. 1973. Pragmatic language testing. *Language Sciences 12*.

Petersen, C. R. and F. A. Cartier. 1975. Some theoretical problems and practical solutions. In R. L. Jones and B. Spolsky, eds. *Testing language proficiency*. Arlington, Va.: Center for Applied Linguistics.

Stevenson, D. K. 1974. A preliminary investigation of construct validity and the Test of English as a Foreign Language. Ph.D. dissertation. Albuquerque, N.M.: University of New Mexico.

Upshur, J. A. 1976. Discussion of J. W. Oller and K. Perkins, A program for language testing research. In H. D. Brown, ed. *Papers in second language acquisition*. Ann Arbor, Mich.: Research Club in Language Learning, University of Michigan.

Section I

General Topics

Classification of Oral Proficiency Tests

Harold S. Madsen and
Randall L. Jones
Brigham Young University

Abstract. A recently conducted survey has disclosed that during the past few years there has been a significant increase in the development of speaking tests. Basic considerations in preparing a speaking test include the purpose for its use (e.g., academic or vocational), the background of the examinee (e.g., age, proficiency level, language experience), the criteria selected (e.g., linguistic or communicative), and the scoring procedure.

In this study we have isolated over two dozen elicitation techniques, which range from measures of conversational spontaneity to measurement of specific linguistic subskills. At one end of the spectrum are informal, open-ended techniques used in some interviews. Slightly more control is available in the pseudo-communicative variety, such as role play. Still more structured are connected discourse techniques, such as reading a prose passage aloud, and controlled responses, like those requiring description of a picture.

A typical composite oral proficiency test for adults would incorporate several elicitation techniques and discrete scoring. It would be administered live, one-on-one, in about ten minutes to a literate examinee. Evaluation of an oral proficiency exam is somewhat relative, depending primarily on its intended use.

Introduction

During the past few decades oral language testing has had a great deal in common with physical fitness. Everyone thinks that it is a wonderful idea, but few people have taken time to do anything about it. During the prime period of audio-lingual methodology, for example, the teaching of oral production was the principal classroom objective, but the testing of oral proficiency was almost unknown. Anyone searching bibliographies that deal with language teaching in the 1950s and early 1960s will come away with precious little information about the testing of speaking.

Matters have apparently changed considerably during recent years. As our contribution to the colloquium on the validation of oral proficiency tests, we were asked to attempt a classification of existing oral language examinations. Our initial reaction was, "What is there to classify?" The FSI is well known, as is the *Ilyin oral interview*. And each of us is aware of a handful of other procedures, but certainly, we assumed, we are dealing with no more than a dozen or so at the most. Our assumption proved very wrong. At Adrian Palmer's request, tests were sent to us. We scoured our own files as well as the journals and requested an ERIC computer search. It soon became obvious that there was far more material to deal with than we had previously thought. When we reached the point that we had approximately one hundred exams, we decided to end the search and begin the classification. For some tests, we have very complete documentation; for others we are only aware of their existence. We expect that the collective pool of knowledge at the colloquium will shed much more light on these and other oral tests.

Reliability and Validity

Before getting into the details of classification, it is appropriate to discuss briefly two important concerns that relate to oral proficiency testing, viz., reliability and validity.

One of the major reasons that so many language teachers have avoided testing oral proficiency directly is due to the apparent problem of reliability. Indeed, as Spolsky has pointed out, the "psychometric-structural" movement in the 1950s was in part a reaction against the subjective testing methods that have been used in the language classroom (Spolsky, 1975). Even though it seemed obvious that an examinee needed to speak if his speaking proficiency were to be tested, it seemed equally obvious to many that there was no consistent method of quantifying the information that is contained in the act of speaking. Because objective tests and other paper-and-pencil tests are so appealing, they have become standard in most language programs. Fortunately, it has occurred to some to actually measure the reliability of oral tests. We now have good empirical evidence that accurate and consistent judgments about speaking proficiency can be made (Clark, 1978a).

The question of validity is another matter altogether. Because a face-to-face oral test so closely approximates a real-life situation, it obviously has high face validity. Most people have thus simply assumed that an oral test is generally a valid instrument. But there are serious potential problems with content validity. On the one hand, much of the data generated in an oral test is superfluous or redundant, while on the other hand there are many important linguistic structures that are not produced. Because the language is random, it is not a good sample of what is taught in the classroom or what is considered to be a minimal standard of proficiency at any particular level. Attempts can, of course, be made

to elicit certain structures or lexical items, but the test then becomes less natural. A compromise between naturalness and efficiency must be made. The tests that we have seen range all the way from discrete-item tests of vocabulary and grammar to general conversation tests. In many cases the validity has yet to be established.

Considerations

Most oral language tests are designed with some specific purpose in mind. No one test can be universally valid, regardless of how it may perform for a given task. Tests can therefore be classified according to the conditions imposed on them. The considerations that are discussed here are not necessarily listed in order of importance.

Academic and nonacademic differences

Most of us are involved in language teaching at an academic institution, and our testing program is very much a part of — or even an adjunct to — our teaching program. Testing is important in determining grades and placement, motivating students, and providing diagnostic feedback for teacher and student. We are not often concerned about how proficient our students are in comparison to students in other parts of the country, or foreign service officers, or other United States citizens working abroad. Most of us are also seldom concerned about how the oral proficiency of our students relates to a particular occupational need, e.g., determining the nature of a patient's problem at a medical clinic or explaining legal rights to a person who has been arrested. But there are many people out there in the real world whose interest in language testing relates directly to job-oriented tasks. They are only interested in knowing how well an examinee will perform on the job. It seems obvious, then, that the design of an oral test should depend partly on the objectives it is intended to meet, and that these objectives differ to some degree between the academic and nonacademic worlds.

Although we generally think of a speaking test as an integrative test of oral production, some testing techniques (particularly in academic settings) narrow the focus to very discrete elements of the language. For example, an examinee might be asked to say the word for an item in a picture (vocabulary), or asked to say a word that is written on a card (pronunciation), or asked to give the past tense form of a verb (inflection). These techniques do not necessarily indicate how well the examinee can carry on a conversation in the language, but they do permit the examiner to focus in on specific items.

Such discrete-item approaches have three advantages. First, they are very efficient. They require a short response; thus much information is obtained in a

relatively short time. Second, they provide very useful diagnostic information. For any particular item the response is either correct or incorrect; thus it is quite apparent where the speaker's strengths and weaknesses lie. Because the speaker is forced to respond to a specific item, he cannot evade it as is often possible in an oral interview. Finally, such items are very easy to score. Because the response is either right or wrong, there is little problem in quantifying the performance of the examinee. The major disadvantage of this type of technique should also be mentioned, viz., that much of the same information can be obtained by a paper-and-pencil test at a fraction of the cost.

Although the purpose of most speaking tests is to measure proficiency or achievement, oral testing can also be useful for diagnostic evaluation and for research. In fact, some of the tests in our list were designed specifically for obtaining research data. There is at the present time a great deal of interest in studying the order of acquisition of linguistic elements among second language learners. Data must be elicited from subjects very much as it is in an oral test. The primary difference is that for research there is usually no need to determine an overall score for the performance.

Level of language proficiency

The proficiency level of the examinee is an important consideration in designing or selecting an oral test. Although tests such as the FSI interview are intended to measure the entire spectrum of proficiency, the techniques employed must differ depending on whether the examinee is at the beginning, intermediate, or advanced level. (These three levels refer to absolute proficiency, not simply levels of achievement in a university language program.) For example, an examinee at the beginning level might have difficulty engaging in a sustained conversation, but could perform well in a simple role-playing situation. An examinee at a high level may not be sufficiently challenged by a general conversation, but could demonstrate his ability well if asked to explain his point of view on a complex abstract topic. There is good reason to believe that most oral tests and testing techniques do not discriminate well at the higher levels of proficiency (Jones, 1978).

In some cases the proficiency of a population group may cluster tightly at some point. The testing instrument would then have to be capable of making fine discriminations within a narrow range of proficiency. In other cases the proficiency may be scattered over a wide spectrum. For the latter situation, the instrument should thus be capable of measuring accurately across several levels. One could compare such an oral test with a quality short-wave radio. The tuning mechanism should cover all frequency bands, but it should have a fine-tuning device for discriminating between frequencies that are very close.

learning history of the examinee. Even though we want to believe that our tests measure general proficiency, we can perceive a difference between the examinee who learned the language in the classroom and the one who learned it in a natural setting. (This difference relates closely to Krashen's distinction between language "learning" and language "acquisition.")

Twice during the past few months students at our institution have requested credit by examination for second semester German (German 102). In both cases the students had lived in Germany for extended periods of time. In neither case had they had extensive instruction in the structure of the language. According to our established procedure, they took the final examination for German 102 (a standardized multiple-choice exam) and in addition had a brief oral examination. In the oral exam both of them performed better than even the top students enrolled in the course. On the written exam one scored B and the other B−.

This distressing discrepancy merely points out that the proficiency profile of the classroom learner is often very different than the profile of the natural-setting learner. How this difference should be reflected in a testing program — if at all — is not entirely clear. One should, however, be prepared to expect differences.

Examinee-examiner language backgrounds

The native language homogeneity of a testing population may seem relatively unimportant, but it does have some bearing on a testing situation. In a typical foreign language program, e.g., German in an American university, the majority of students have a common first language. In a typical second language program, e.g., German for foreigners at the Goethe Institute, the opposite is true. The major consideration here is that certain techniques can be employed only if the examiner knows the native language of the examinee. This is almost always true in a foreign language program, but rarely the case in a second language program. An interpreter task, a technique frequently employed at the FSI, would be next to impossible in most ESL programs.

Another important factor involving the native language of the examinee is the effect that obvious native language interference errors have on the examiner. An examiner who understands the native language of the examinee may subconsciously overlook certain errors simply because he is so used to hearing them in the classroom. Errors made by examinees whose native language is not known by the examiner may be scrutinized more carefully.

A minor but certainly not insignificant consideration has to do with minimal prerequisites of the examinee. For example, if a given test presupposes a certain level of proficiency, that fact should be well understood. Testers at the FSI occasionally find themselves in an embarrassing situation when an examinee fresh from college presents himself for an oral interview. In spite of A work in four semesters of the language, the hapless student is not able to even begin a

basic conversation. His perception of his ability to use the language is considerably different from what is measured on the FSI rating scale.

Another prerequisite has to do with the use of other skills. An examinee may be asked to read and summarize a short passage in the foreign language. This is only possible if he is able to read the language. Some language teachers use special symbols to facilitate language use in the classroom. These same symbols can be useful in a testing situation, but only if all examinees are familiar with them.

Procedure

Although a real direct test of oral proficiency would involve observing an examinee using the language in a natural situation, most testing programs cannot afford this luxury. Instead, the tests usually consist of a face-to-face encounter in which all situations are simulated. Using a variety of techniques, the examiner can elicit speech samples to be evaluated. It is also possible for the examiner to interact with several examinees during the same test, and for the examinees to interact with each other. Such a group testing is not only efficient, but it also allows linguistic interaction among peers. Some techniques require no live examiner at all, but rather use printed and recorded stimuli, with all responses recorded for later evaluation. Such techniques do not allow for spontaneous conversation, nor do they provide a very natural setting for communication, but they generally produce consistent results.

Criteria

For many years linguistic criteria were the only ones considered in language testing. A person's ability to communicate was assumed to be related directly to his ability to control the linguistic elements of the language, i.e., pronunciation, grammar, and vocabulary. Later, fluency was added to the list, even though it was not at all certain that everyone agreed on what it meant.

Recently, a rising interest in communicative competence has forced us to examine more closely what else besides linguistic facility contributes to effective communication in a second language. Unfortunately, communicative competence has come to mean many things to many people, and it is not a term that is unambiguously understood among language teachers. But certainly a sensitivity to appropriateness of language and an understanding of nonverbal paralinguistic signals are important. Unfortunately, these additional features pose very difficult problems for testing.

Elicitation Cues

A test item consists of a stimulus and a response. In an oral test the response must by definition be spoken, but the stimulus might be oral, visual, or a

combination of the two. For example, an examiner might show a picture and ask the examinee to identify or explain something in it. Or he might ask the examinee to interpret a gesture or facial expression. Some stimuli ask for a very general response, e.g., "Tell me about this picture," or "How do you like Boston?" Others are more specific, e.g., "What is this?" or "Where do you live?"

Scoring Procedures

When we speak of a test as being either objective or subjective, we are referring not to the test itself, but rather to the procedure for scoring it. In an oral test, validity is very closely related to elicitation procedures, while reliability is more closely related to the scoring procedure. For an oral test it is necessary somehow to translate observations into numerical or verbal scores.

For a discrete-item oral test the score can be determined simply by adding up the number of correct responses. For more integrative tests, two basic approaches can be used: a rating scale or a holistic evaluation. A rating scale is usually accompanied by definitions of the performance at various levels. A number is assigned for each factor, and a total score is determined by adding up the points. Some scales use only linguistic criteria (e.g., FSI; Clark, 1972); others include additional factors (e.g., Bartz, 1974; Schulz, 1974). Most FSI testers are trained using the rating scale, but later arrive at a score using a holistic evaluation.

The ideal oral examiner is a trained specialist who is in no way biased toward any of the examinees. Unfortunately, such ideal conditions rarely exist anywhere. Teachers usually have to test their students, and in some tests the students themselves participate in evaluating their peers. Because the accuracy of an evaluation is directly related to the training of the evaluator, it is vital that the criteria and scoring procedure be clearly defined and understood. Where possible, it is useful to employ more than one evaluator. This provides a built-in check for consistency, and allows the scorers to discuss their decisions in cases of discrepancies. At the FSI the score is determined by the linguist, with the native speaker providing a control. At the CIA each of the two testers makes an independent rating. The two ratings are checked to make certain that they are within a defined tolerance range.

Where feasible, it is most efficient to determine the score of a test immediately after it has been administered. If a recording of the test is used, it provides the evaluator an opportunity to review the test carefully, but it can also obscure the impression that one gains from observing a live spontaneous situation. If a recording of a test is the only basis for judgment, the evaluator misses all of the important paralinguistic signals.

Oral Testing Techniques

While we have isolated over two dozen techniques in oral proficiency tests, these can be grouped into a few broad categories reflecting elicitation strategy and the focus of the evaluation. At one end of the spectrum are question types designed to generate communicative language; at the other end, techniques to facilitate discrete measurement or evaluation of specific subskills.

Communicative discourse

The most frequently used approach in the 60 tests analyzed for this study is a direct measure of speaking ability through conversation. The usual technique is Question and Answer. This form varies from fixed questions ("What is your name?" / "My name is Mohammed Nassr.") to the rather spontaneous ("What made you decide to become a nurse?" / "Well, my mother was a nurse, and . . ."). In addition to this approach, a few tests incorporate the complementary Statement-Response form ("I'm sorry you had to wait so long." / "That's quite all right." Madsen and Taylor, 1971). Very few, however, incorporate ambiguities, obscured cues, faulty information, and the like in order to prompt self-initiated responses on the part of the examinee ("Take this to the other room, please." / "Pardon me. Which room is that?"). But to promote interaction that is as genuine as possible, some test writers specify that "free conversation" is to be conducted on some topic.

Occasionally naturalness is also sought for by removing the examiner from direct conversation with the examinee, yet retaining the element of human interaction. One such device is the Dyad, where a student exchanges information with a peer, in activities ranging from evaluating each other's oral reading to problem solving (Findley, 1977). Another is Group Evaluation of five to seven students (Folland and Robertson, 1976). A film or tape can provide a topic of common acquaintance. The group then discusses the topic at hand while one or more judges evaluate individual responses. When multiple examiners are used, each can evaluate a separate language feature for all participants; or each can make a total evaluation of one student.

Pseudo-communicative discourse

To provide somewhat more control over the language produced by the examinee and still maintain a communicative form, some testers prefer a slightly less direct oral examination procedure. One technique is Role Play. Usually a variety of situations are provided, and the examiner selects one at random. He may carry out a fixed role, with the examinee interacting spontaneously. In a classroom setting, two or more students can participate, the teacher-rater simply acting as an observer (Valette, 1977). Subjects can range from declining a date to changing travel arrangements.

Another form of pseudo-communication is the Directed Request, a task not uncommon in the everyday world: "Would you please ask that man if we could look at his telephone directory for a moment?" / "Excuse me. Can we use your directory for a few minutes?" Yet another is the Interpreter Task, frequently included in the FSI interview. The examiner assumes the position of a monolingual who speaks only the native language of the examinee. The former reportedly needs to communicate with a second party who speaks only the language being evaluated. The examinee, therefore, finds it necessary to engage in two-way translation: native language to foreign language, and foreign language to native language.

Connected discourse

Of the several ways that connected discourse can be generated, some — such as giving a talk — approximate communication in real life, while others — such as providing a narration from picture cues — are less natural, or, as Clark has indicated, "indirect" (1978b). Yet each maintains that flow of language generally felt to typify real communication.

Aside from conversation techniques, the most popular means of generating connected discourse is simply to have the candidate read a passage aloud. This obviates the necessity of finding a suitable topic; it standardizes the output and generates precisely the language desired. But the Read Aloud technique has some obvious limitations: It cannot be used with children who have not yet learned to read, or with candidates whose oracy substantially exceeds their literacy. And while it provides a measure of pronunciation, it hardly measures communicative skills such as fluency and appropriateness. People with equal proficiency in speaking often vary significantly in their ability to read aloud from a script.

Other connected-discourse techniques are more cognitive. One exam that utilizes a reading passage requires the student to *explain* what he has read (*Spoken English for industry and commerce*, n.d.). Several tests require candidates to retell a story that is presented to them orally. Circumventing the memory problem associated with the Retold Story is the Narrative from Pictures approach, which has candidates create a narrative from ideographs or multiple sketches. Section 6 of the ARELS exam requires students to select one of twelve topics a few days before the test is administered and then as part of the exam to speak for 60 seconds on the subject without notes. Normally the presentation is extemporaneous. On one test the ESL student hears a question in his native language and then in English. An easy question requires a 15-second response ("Where have you taught school and where do you now teach?"); a more difficult question may take up to 45 seconds ("Describe a typical day at your school." Rand, 1968). Yet another question requests short monologues including such matters as apologies, excuses, invitations, complaints, etc. ("You are in a restaurant. The plate you are given by the waiter is dirty." Levenston, 1973).

Two additional types of connected-discourse techniques are Explanation and Description. The former could include an item such as "Explain how American children celebrate Halloween." The latter might incorporate an item such as "Describe a guitar." In brief, questions in this area vary in degree of control as well as difficulty, but all require varying amounts of connected speech.

Controlled response

There is a continuum from the techniques we have just discussed and the mechanical, discrete items found on some oral exams. Bridging the two extremes are open-ended items that permit flexibility in response. One rather popular approach is the Visual + Description item. This can consist of an extended (possibly rambling) description of the items or activities represented in the sketch; or it might constitute a one-sentence explanation of a simple line drawing. An advanced student might be required to explain a technical graph (*The English for business test*, n.d.); a bilingual child might be asked to describe an object he can see (Evans, n.d.). In a problem-solving situation, the student might have to describe one picture from a set so that a native speaker can identify the sketch in question. In a Visual + Student Question item, the student attempts to identify one particular picture by asking questions. An example of this takes the following form (Palmer, 1971):

Student	*Examiner*
1. Is the man sitting on the floor?	NO
2. Is he sitting on the chair?	YES
3. It's number 1.	CORRECT

(Sketches include a man sitting on a chair, a man sitting on the floor near a chair, a man standing on a chair, and a man standing on the floor near a chair.)

A number of more restricted techniques are also available. One of the most popular is Elicited Imitation. This features the control of Reading Aloud, but it is available to children unable to read, and to mature beginning students. Because of the memory factor, seldom is more than one sentence read at a time, and disconnected sentences or even single words may be elicited. One test presents the sentences orally and then has the student read them aloud (Pimsleur, 1967). Another presents large enough oral chunks that short-term memory is exceeded and the examinee's underlying grammatical competence is thereby evaluated (Swain et al., n.d.)

The Directed Response is likewise quite controlled. A rather easy item might take this form: "Tell me that you like fish." / "I like fish." An advanced version has appeared thus: "An urgent letter your secretary's typed is full of mistakes: without offending her persuade her to do it again." / "There are one

or two small errors in this letter; do you think you could perhaps do it again?" (*ARELS oral examinations*, n.d.)

Directed Affect also involves brief instruction followed by a short response. But instead of syntactic or lexical adjustments, the focus is on tone or affect. One test has the respondent say "hello" first with a single affect, such as "sadness," to three affects, such as "likes me, is serious, is younger than I am" (Heinberg, et al., 1970). Another test uses different utterances and different affects for every question (Palmer, 1974):

> "Did you notice how high the water was?"
> (a) worried (b) matter of fact
> (Note: The cues "worried" and "matter-of-fact" are printed in the native language, Thai.)

Linguistic skill

Oral tests that attempt to measure specific linguistic skills range from the communicative to the mechanical. For instance, the *Bilingual syntax measure* (Burt et al., 1975) utilizes natural exchanges of conversation between a child and an examiner in relation to a series of pictures, yet the scoring focuses exclusively on grammar. There is a tendency, however, for tests that quantify linguistic accuracy or complexity to opt for controlled responses. Acting as interpreter is one means of evaluating mastery of syntax. Sentence Completion is another: "I live . . ." / "I live in Chicago." Still another is Grammatical Manipulation: "Make a question out of this sentence: She's tall." / "Is she tall?"

In addition to syntax, phonology can be evaluated through such devices as Elicited Imitation (mimicry of spoken words and phrases) or Reading Aloud — very popular techniques. Several tests even use the Bipolar Response, with its minimal oral utterance. For instance, upon hearing a minimal pair ("sit-seat"), the candidate simply says "Different" to indicate that the two words are not identical.

Vocabulary receives a surprising emphasis in contemporary oral tests. One technique is Directed Translation: "Was bedeutet 'Buch' auf Englisch?" / "Book." The single most frequently used method is Picture-cued Vocabulary. Such items range from individual sketches of an object or actual realia to a complex drawing of buildings and streets intended to elicit "city." Other approaches include Oral Cloze and Synonym-Antonym production. The latter is illustrated in a bilingual test requiring the student to provide opposite expressions for stimulus words (*Dos Amigos verbal language scales*, n.d.)

Though not a linguistic subskill, listening proficiency is also evaluated separately in many "oral production" tests. One of the most common ways is by

linking an oral cue with a printed multiple-choice response. Consider an "appropriate response" example:

"How far is it to Boston?"
(A) No, not far.
(B) North of New York.
(C) About 50 miles.

Another frequently used technique, especially with children, is the "pure" response consisting of pointing at a picture that best matches the stimulus cue. A third procedure is TPR (total physical response): "Put the pencil on top of the book." / (Student carries out the request). Finally, an unusual mode for a speaking test is an optional native-language response to demonstrate that comprehension has occurred (De Avila and Duncan, 1977).

A Profile of Oral Tests

A look at nearly a hundred oral proficiency tests reveals some interesting contemporary trends. For one thing, this sizable number of exams refutes the commonly held notion that nothing is being done in oral testing. A substantial proportion of contemporary commercial tests were developed in Great Britain. The bulk of American commercial oral tests were designed for bilingual purposes. Among the most prominent American general proficiency ESL batteries — TOEFL, Michigan's MTELP, the CELT, and ALIGU — only the ALIGU provides even an optional oral test, and it is seldom used. In brief, most oral exams — particularly in the United States — have been created independent of existing batteries. This is reflected in those most widely recognized in American ESL circles: the FSI (*Foreign Service Institute oral interview test*, n.d.), the *Ilyin oral interview*, and the *Bilingual syntax measure* (Burt et al. 1975).

An analysis of approximately five dozen contemporary oral tests reveals that the vast majority incorporate subtests and multiple elicitation techniques. And without abandoning their interest in integrative examinations, test makers evidence a strong interest in approaches that are quantifiable (e.g., number of responses in 30 seconds, exact word criteria in elicited imitation, readily identifiable answers to picture-cued questions). Live interaction is still preferred, even in nonbilingual commercial tests (approximately 60 percent), with two-thirds of the experimental tests and almost 90 percent of the bilingual tests utilizing live examination procedures. While about half of the commercial and experimental tests utilize printed stimuli or instructions, virtually no bilingual tests do so. Fewer than a third of all oral tests involve a taped recording of the examinee. Several British tests (but no American test surveyed) provide separate examination forms for different grade or ability levels. The time required for oral tests ranges from under five minutes to a high of fifty minutes, with a median time of ten minutes.

Virtually all bilingual tests and the majority of experimental exams provide specific measurement of one or more linguistic subskills — syntax, phonology, or lexis. Nonbilingual commercial tests are somewhat less inclined to do so. In oral tests measuring such subskills, 50 percent more quantifying is done of structural proficiency than of either phonological or lexical proficiency. With regard to age level, nearly all bilingual tests are designed to be used with children, although some can be employed equally well with adults. The bulk of nonbilingual oral tests, on the other hand, are aimed at the post-elementary school audience.

A typical composite oral proficiency test for adults would incorporate at least two elicitation techniques and discrete scoring. Syntactic control would be evaluated in one subsection. The test would be administered one-on-one with a live examiner, but would not be tape-recorded. Not part of a larger battery, this test would require ten minutes to administer. Examinees would be literate in the target language, but the examiner would not need to have sophisticated linguistic skills in order to administer and score the test.

While our composite exam may be typical, it is not necessarily "ideal." The many different oral test formats do not so much represent confusion as, rather, attempts to meet the special evaluation needs referred to earlier. Tests prepared for young children obviously avoid printed cues, as do the occasional exams prepared for illiterate adults. Those used as research instruments in language acquisition (e.g., Fathman, 1975) may rely on an accurate assessment of syntactic mastery, while an evaluation of communicative competence might justify a look at problem solving or even interaction in a group.

Thus in selecting appropriate instruments for a convergent-divergent validation study, an important consideration might well be the availability of a parallel test form (to the FSI, for instance) in another modality. In short, the seeming plethora of oral proficiency examinations can enable the user to select or design an instrument more suitable than ever before to his particular testing requirements.

REFERENCES

ARELS oral examination (ARELS). n.d. London: The Examinations Trust of the Association of Recognized English Language Schools.

Bartz, Walter H. 1974. A study of the relationship of certain factors with the ability to communicate in a second language (German) for the development of communicative competence. Ph. D. dissertation. The Ohio State University.

Burt, Marina K., Heidi C. Dulay, and Eduardo Hernandez-Ch. 1975. *Bilingual syntax measure* (BSM). New York: Psychological Corporation. (Also: San Francisco: Harcourt Brace Jovanovich.)

Clark, John L. D. 1972. *Foreign language testing: theory and practice*. Philadelphia: Center for Curriculum Development.

_____. 1978a. *Direct testing of speaking proficiency*. Princeton: Educational Testing Service.

_____. 1978b. Psychometric considerations in language testing. In Bernard Spolsky, ed. *Approaches to language testing*. Papers in applied linguistics (Advances in language testing series: 2). Arlington, Va.: Center for Applied Linguistics.

De Avila, Edward A. and Sharon E. Duncan. 1977. Language assessment scales (LAS I). n.p.: Linguametrics Group, Inc.

Dos Amigos verbal language scales (DAVLS), n.d. San Rafael Cal.: Academic Therapy Publications. (English/Spanish)

The English for business test (the ELTDU test; the Bellcrest test). n.d. Colchester, Essex, England: English Language Teaching Development Unit of Oxford University Press.

Evans, Joyce. n.d. *Spanish/English language performance screening*. Austin, Texas: Southwest Educational Development Laboratory.

Fathman, Ann. 1975. The relationship between age and second language productive ability. *Language Learning 25*: 245-253. (See also: S. Krashen, S.V. Sferlazza, and A. Fathman. 1976. Adult performance on the SLOPE test: more evidence for a natural sequence in adult language acquisition. *Language Learning 26*: 145-151.)

Findley, Charles A. 1977. Dyadic task-oriented communication exercises for teaching and testing in the elementary ESL class. ED 145 692.

Folland, David and David Robertson. 1976. Toward objectivity in group oral testing. *English Language Teaching Journal 30*: 156-167.

Foreign Service Institute oral interview test (FSI). n.d. Washington, D.C.: Foreign Service Institute.

Heinberg, Paul, Burton Byers, Arthur Coladarci, and L. S. Harms. 1970. *Hawaii communication test* (HCT). Honolulu: University of Hawaii.

Ilyin, Donna. 1976. *Ilyin oral interview*. Rowley, Mass.: Newbury House.

Jones, Randall L. 1977. Testing: a vital connection. In June K. Phillips, ed. *The language connection: from the classroom to the world*. Skokie, Ill.: National Textbook Company. 237-265.

_____. 1978. Interview techniques and scoring criteria at the higher proficiency levels. In Clark, 1978a: 89-102.

Levenston, E. A. 1973. *Test of oral proficiency of adults*. Toronto, Canada: Ontario Institute for Studies in Education.

Madsen, Harold S. and James S. Taylor. 1971. *WIN test of oral proficiency* (WIN TOP). Sacramento: California State Department of Education.

Palmer, Adrian S. 1971. *Oral communication test* (COMTEST). Bangkok, Thailand: Thammasat University.

_____. 1974. *Oral communication test for speakers of Thai*. Khon Kaen, Thailand: English Department, Khon Kaen University.

Palmer, Adrian S. and Jack Upshur. 1971. *Oral production test* (PROTEST). Bangkok, Thailand: Thammasat University. (Form A)

Pimsleur, Paul. 1967. *Pimsleur modern foreign language proficiency tests*. New York, N.Y.: The Psychological Corporation.

Rand, Earl. 1968. *A short test of oral English proficiency*. Austin, Texas: University of Texas and Taiwan Provincial Normal University.

Schulz, Renate A. 1974. Discrete-point versus simulated communication testing: a study of the effect of two methods of testing on the development of communicative competence in beginning French classes. Ph.D. dissertation. The Ohio State University.

Spoken English for industry and commerce (the London Chamber of Commerce and Industry tests). n.d. Sidcup, Kent, England: The London Chamber of Commerce and Industry.

Spolsky, Bernard. 1975. Language testing: art or science? Paper presented at the Fourth AILA World Congress in Stuttgart, Germany.

Swain, Merrill, G. Dumas, and N. Naimen. n.d. Alternatives to spontaneous speech. EDRS ED 123 872.

Valette, Rebecca M. 1977. *Modern language testing*, 2nd ed. New York: Harcourt Brace Jovanovich.

*A Theoretical Framework for Communicative Competence**

Michael Canale and Merrill Swain
The Ontario Institute for Studies in Education

Abstract. This paper briefly outlines the contents and boundaries of three areas of competence, or systems of knowledge, that are to be minimally included in a theory of communicative competence: grammatical competence, sociolinguistic competence, and strategic competence. Grammatical competence is concerned with the rules of sentence grammar and sentence grammar semantics. Sociolinguistic competence includes sociocultural rules for determining the social meaning and appropriateness of a single sentence or utterance and discourse rules for determining the cohesion and coherence of groups of utterances. Strategic competence is composed of verbal and nonverbal communicative strategies that are used to compensate for breakdowns in communication due to performance factors or to insufficient grammatical or sociolinguistic competence. It is suggested that the value of such a theoretical framework for second language learning is that it provides a clear initial statement, or construct, of communicative competence. Such a statement is helpful not only for the purposes of second language teaching but also for those of second language testing.

Introduction

During the past eight months we have been working to determine the feasibility and practicality of measuring the 'communicative competence' of students enrolled in general French as a second language programs in elementary and

*The research reported here was carried out on the project 'French as a second language: Ontario assessment instrument pool' and was funded under contract by the Ministry of Education, Ontario. We gratefully acknowledge this support. We also wish to express our thanks to Andrew Cohen, Alan Davies, Bruce Fraser, Peter Groot, Randall Jones, Adrian Palmer, and Joel Walters for helpful discussion of the ideas presented here. Of course, none of these people is responsible for the opinions expressed here or for any form of error.

secondary schools in Ontario. In Canale and Swain (1979) we argued for a theory of communicative competence that minimally includes three main competencies, or systems of knowledge: grammatical competence, sociolinguistic competence, and strategic competence. The purpose of this paper is to briefly outline the contents and boundaries of each of these areas of competence.

Orientation

Following Morrow (1977), we understand communication to be interaction based, to involve unpredictability and creativity, to take place in a discourse and sociocultural context, to be purposive behavior, to be carried out under performance constraints, to involve use of authentic (as opposed to textbook contrived) language, and to be judged as successful or not on the basis of behavioral outcomes. Furthermore, communication will be understood to involve verbal and nonverbal symbols, oral and written modes, and production and comprehension.

We will assume that a theory of communicative competence interacts (in as yet unspecified ways) with a theory of human action and with other systems of human knowledge (e.g., world knowledge). We will assume further that communicative competence, or more precisely its interaction with other systems of knowledge, is observable indirectly in actual communicative performance. These assumptions have been discussed in Canale and Swain (1979).

The theoretical framework that we propose is intended to be applied to second language teaching and testing in line with the communicative approach outlined in Canale and Swain (1979). This approach is an integrative one in which emphasis is on preparing second language learners to exploit those grammatical features of the second language that are selected on the basis of (among other criteria) their grammatical and cognitive complexity, transparency with respect to communicative function, probability of use by native speakers, generalizability to different communicative functions and contexts, and relevance to the learners' communicative needs in the second language. To do this, they initially draw on aspects of the sociolinguistic competence and strategic competence they have acquired through experience in communicative use of their first or dominant language. Our thinking in developing this theoretical framework and communicative approach owes much to the scholarship of Allen and Widdowson (1975), Halliday (1970), Hymes (1967; 1968), Johnson (1977), Morrow (1977), Stern (1978), Wilkins (1976), and Widdowson (1978).

Components of Communicative Competence

Grammatical competence

This type of competence will be understood to include knowledge of lexical items and of rules of morphology, syntax, semantics, and phonology. It is not

clear that any particular theory of grammar can at present be selected over others to characterize this grammatical competence, nor in what ways a theory of grammar is directly relevant for second language pedagogy (cf. Chomsky, 1973, on this point), although the interface between the two has been addressed in recent work on pedagogical grammars (cf. Allen and Widdowson, 1975, for example). Nonetheless, grammatical competence will be an important concern for any communicative approach whose goals include providing learners with the knowledge of how to determine and express accurately the literal meaning of utterances.

Sociolinguistic competence

This component is made up of two sets of rules: sociocultural rules of use and rules of discourse.

Sociocultural rules of use. These rules will specify the ways in which utterances are produced and understood *appropriately* with respect to the components of communicative events outlined by Hymes (1967; 1968). It should be emphasized that it is not clear that all of the components Hymes proposed are necessary to account for the appropriateness of utterances or that these are the only components that need to be considered. Knowledge of these rules will be crucial in interpreting utterances when there is a low level of transparency between the literal meaning of the utterance and the speaker's intention, i.e., the social meaning or value of the utterance.

Rules of discourse. Until more clear-cut theoretical statements about rules of discourse emerge, it is perhaps most useful to think of these rules in terms of the cohesion (i.e., grammatical links) and coherence (i.e., appropriate combination of communicative functions) of groups of utterances (cf. Halliday and Hasan, 1976, and Widdowson, 1978, for discussion). It is not altogether clear to us how rules of discourse will differ from grammatical rules (with respect to cohesion) and sociocultural rules (with respect to coherence). However, the focus of rules of discourse is the combination of utterances, not the grammatical well-formedness nor the social meaning or appropriateness of a single utterance. Also, rules of discourse will presumably make reference to notions such as topic and comment (in the strict linguistic sense of these terms) whereas grammatical rules and sociocultural rules will not necessarily do so.

Strategic competence

This component will be made up of verbal and nonverbal communicative strategies that may be called into action to compensate for breakdowns in communication due to performance variables or insufficient competence. Such strategies will be of two main types: those that relate primarily to grammatical competence (e.g., how to paraphrase grammatical forms that one has not mastered

or cannot recall momentarily) and those that relate more to sociolinguistic competence (e.g., various floor-holding strategies, how to address strangers when unsure of their social status). We know of very little work in this area (though see work by Duncan, 1973; Fröhlich and Bialystok, in progress; Tarone, Cohen, and Dumas, 1976; as well as discussion by Candlin, 1978; Morrow, 1977; Stern, 1978; and Walters, 1978). Knowledge of how to use such strategies may be particularly helpful at the beginning stages of second language learning. Furthermore, as Stern (1978) has pointed out, such 'coping' strategies are most likely to be acquired through experience in real-life communication situations but not through classroom practice that involves no meaningful communication.

Probability Rules of Occurrence

Within each of the three components of communicative competence that we have identified, we assume there will be a subcomponent of probability rules of occurrence. These rules will attempt to characterize the 'redundancy aspect of language' (Spolsky, 1968), i.e., the knowledge of relative frequencies of occurrence that a native speaker has with respect to grammatical competence (e.g., sequences of words in an utterance), sociolinguistic competence (e.g., sequences of utterances in a discourse), and strategic competence (e.g., commonly used floor-holding strategies). Proposals for the formal expression of such rules are discussed by Labov (1972), where it is claimed that various features of the sociolinguistic and grammatical contexts combine to condition the frequency of use of a given rule of grammar. The importance of such rules for communicative competence has been stressed by Hymes (1972) and Jakobovits (1970) and suggested in the work of Levenston (1975), Morrow (1977), and Wilkins (1978). Related to the discussion of these rules is the proposal that authentic texts be used in the second language classroom from the very beginning (cf. Morrow, 1977, for discussion). Although much work remains to be done on the form of such probability rules and the manner in which they are to be acquired, the second language learner cannot be expected to have achieved a sufficient level of communicative competence in the second language, in our opinion, if no knowledge of probability of occurrence is developed in the three areas of communicative competence.

Conclusion

In proposing such a theoretical framework for communicative competence, it is expected that the classification of communication skills suggested by Munby (1978) will serve as an initial indication of the types of operations, subskills, and features that are involved in successful communication. Certainly there will be modifications to this classification scheme (e.g., the addition of skills relating to strategic competence), just as there will no doubt be modifications to the general

theoretical framework outlined briefly here. It is hoped that such a framework will help to establish a clear statement of the content and boundaries of communicative competence — one that will lead to more useful and effective second language teaching, and allow more valid and reliable measurement of second language communication skills.

REFERENCES

Allen, J. P. B. and H. G. Widdowson. 1975. Grammar and language teaching. In J. P. B. Allen and S. P. Corder, eds. *The Edinburgh course in applied linguistics*, Vol. 2. London: Oxford University Press.

Canale, M. and M. Swain. 1979. *Theoretical bases of communicative approaches to second language teaching and testing*. Toronto, Canada: The Ministry of Education, Ontario. (To appear in revised form in *Applied Linguistics 1*, 1.).

Candlin, C. N. 1978. Discoursal patterning and the equalising of integrative opportunity. Paper read at the Conference on English as an International and Intranational Language, The East-West Center, Hawaii, April 1978.

Chomsky, N. 1973. Linguistic theory. In J. W. Oller, Jr., and J. C. Richards, eds. *Focus on the learner: pragmatic perspectives for the language teacher*. Rowley, Mass.: Newbury House.

Duncan, H. 1973. Towards a grammar for dyadic conversation. *Semiotica 9*, 1.

Fröhlich, M. and E. Bialystok. In progress. Inferencing strategies for communication. (Work in the Modern Language Center at the Ontario Institute for Studies in Education.)

Halliday, M.A.K. 1970. Language structure and language function. In J. Lyons, ed. *New horizons in linguistics*. Harmondsworth, England: Penguin Books.

Halliday, M.A.K. and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Hymes, D. 1967. Models of the interaction of language and social setting. *Journal of Social Issues 23*, 2: 8-28.

_____. 1968. The ethnography of speaking. In J. Fishman, ed. *Readings in the sociology of language*. The Hague: Mouton.

_____. 1972. On communicative competence. In J. B. Pride and J. Holmes, eds. *Sociolinguistics*. Harmondsworth, England: Penguin Books.

Jakobovits, L. A. 1970. *Foreign language learning*. Rowley, Mass.: Newbury House.

Johnson, K. 1977. The adoption of functional syllabuses for general language teaching courses. *Canadian Modern Language Review 33*, 5: 667-680.

Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Levenston, E. A. 1975. Aspects of testing the oral proficiency of adult immigrants to Canada. In L. Palmer and B. Spolsky, eds. *Papers on language testing 1967-1974*. Washington, D.C.: TESOL.

Morrow, K. E. 1977. *Techniques of evaluation for a notional syllabus*. Reading: Centre for Applied Language Studies, University of Reading. (Study commissioned by the Royal Society of Arts.)

Munby, J. 1978. *Communicative syllabus design*. Cambridge: Cambridge University Press.

Spolsky, B. 1968. Language testing — the problem of validation. *TESOL Quarterly 2*: 88-94.

Stern, H. H. 1978. The formal-functional distinction in language pedagogy: a conceptual clarification. Paper read at the 5th AILA Congress, Montreal, August.

Tarone, E., A. D. Cohen and G. Dumas. 1976. A closer look at some interlanguage terminology: a framework for communication strategies. *Working Papers on Bilingualism 9*.

Walters, J. 1978. Social factors in the acquisition of a second language. Paper read at the 5th AILA Congress, Montreal, August.

Widdowson, H. G. 1978. *Teaching language as communication*. London: Oxford University Press.

Wilkins, D. A. 1976. *Notional syllabuses*. London: Oxford University Press.

_____. 1978. Approaches to syllabus design: communicative, functional or notional. In K. Johnson and K. Morrow, eds. *Functional materials and the classroom teacher: some background issues*. Reading: Centre for Applied Language Studies, University of Reading.

*Beyond Faith and Face Validity: The Multitrait-Multimethod Matrix and the Convergent and Discriminant Validity of Oral Proficiency Tests**

Douglas K. Stevenson
Universität Essen

Abstract. Recently, there has been a renewed international interest in direct oral proficiency measures such as the oral interview. This strong interest has been matched by a growing awareness among some language testing specialists that all proficiency tests must be subjected to construct validation. Unfortunately, the greatest appeal of oral interviews to the technically untrained language teacher rests with their high face validity. This appeal tends to cloud and confuse the need to validate these tests. As a result, although oral interviews are becoming more and more popular among language teachers and testers, this popularity far outruns any technically demonstrated validation, whether content, criterion-related, or construct. In this paper these basic concerns and needs are brought together and made explicit. The climate of validation that supports or hinders the construct validation of oral proficiency tests is described, basic definitions are clarified, and the logic of validation that demands the construct validation of oral proficiency measures is presented and defended. After having made these primary considerations explicit, one central approach to construct validation, convergent and discriminant validation by the multitrait-multimethod matrix, is argued to be the most appropriate for language proficiency tests such as the oral interview. The importance of viewing tests as trait-method units is stressed, with its relevance to language testing theory. A central theme throughout the paper is the interdependencies of language aspects and testing theory in language testing. It is maintained that the strong tendency for language testers to

*This is a considerably shortened version of the paper presented at the 1979 Colloquium on the Validation of Oral Proficiency Tests. The author is grateful to the Deutsche Forschungsgemeinschaft for its support in the preparation and delivery of this paper.

be preoccupied by the language aspect can seriously impair the validation of language tests. Because oral interviews are being used more and more to make decisions, this concern is of more than theoretical interest.

Approaches

We are well aware in language testing that "all the theoretical problems . . . are likely to be present in a concentrated form when trying to measure performance in a spoken language" (Perren, 1968: 108). Similarly, we recognize that a preoccupation with the concept of speaking ability, and how one teaches, learns, or acquires it, is the hallmark of modern language pedagogy. As a construct involving various constitutive or operational definitions, speaking ability is also at the heart of heated debate in modern linguistics and much of modern psychology. However, as Spolsky (1975a) points out, we are much more sensitive to the traffic and trends of linguistics and language pedagogy than we are to those of our other parent discipline, educational and psychological measurement. We therefore tend to overlook the fact that convergent and discriminant validation, as one major approach to construct validation, is an equally complex area. Just as the testing of oral language proficiency is set off from other "proficiencies" because of its complexities, construct validity is set off from the other technical validities (i.e., content and criterion-related) by its "preoccupation with theory, theoretical constructs and scientific inquiry involving the testing of hypothesized relations" (Kerlinger, 1964: 449).

This intersection of complexities dictates a selectivity in any discussion of approaches. My intent in this paper is to emphasize what are two basic considerations in and for such an approach, and to do so from two viewpoints. First, it is too often overlooked that around any measure there exists a "climate of validation". This climate of validation can be defined as the views of validity and validation held by those working with a measure, as well as their needs and expectations for it. The climate of validation surrounding a measure can support, hinder, or effectively deny what I shall call the "spirit of validation." This spirit of validation can be seen as the organized, functionally skeptical, and admittedly somewhat idealized "textbook" view of validity and validation. It is best put forward in the well-known *Standards for educational & psychological tests* (American Psychological Association, 1974; henceforth *Standards*). A serious and critical examination of the climate of validation as it affects (and agrees with) the spirit of validation is a mandatory first step in discussing the convergent and discriminant validation of oral proficiency tests. Such an approach rests upon basic assumptions of terminology, attitudes towards validity and validation, and, not least, "the orientation of the investigator" (Cronbach and Meehl, 1972: 92). Of special interest are the various views held of face validity, content validity, and criterion-related validity, as they pertain to oral proficiency measures.

Second, after some of these considerations and their importance have been made explicit, the central approach to convergent and discriminant validation, that of Campbell and Fiske (1959), is discussed. The emphasis here is on the theoretical assumptions and demands of this approach, but some practical suggestions for the implementation of this approach in a planned large-scale study are also given.

Validation and rationalization

One of educational and psychological measurement's most respected spokesmen, Richard Ebel, has stated (1961: 640) that "validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few." It is the *primary* importance of validity which has most often been the theme of praise, as for instance in Cronbach's (1970: 121) statement that "the quality that most affects the value of a test . . . is its validity," or in Spolsky's (1968a) widely reprinted statement that "the central problem of language testing, as of all testing, is validity."

It is most important to the discussion that follows, and to the success of the planned large-scale study, that this difference between praise and validation be seen, and taken as a starting point. If we are to objectively consider the climate of validation for oral proficiency tests, we must also critically examine what is one of the basic reasons for this colloquium, which is that the popularity of the oral interview as a technique has far outrun its verified technical validity as a measure (of which the FSI Oral Interview is the best example). Moreover, it has escaped from its original in-house, and therefore more closely controlled, use (cf. Wilds, 1975: 35), into the far less controlled and far larger field of education. It has in fact been more praised than validated.

We must begin by accepting validity as the *central* problem, and of *primary* importance. Whether or not we *believe* that "the interview technique as a measure of real-life proficiency" (Clark, 1978b: 225) is *probably* valid is not the point, at least not in the hard-nosed tradition represented by the spirit of validation. The measure purports to reflect an abstract variable, a construct ("real-life speaking proficiency"), however loosely defined. In spite of how much we might believe it reflects such a construct, "rationalization is not construct validation" (Cronbach and Meehl, 1972: 105). In short, our situation closely resembles the one described by Ebel, and this situation is not compatible with the spirit of validation required for construct validation approaches. Because of the primary importance of validity and validation, it is of primary importance that we consider why such interview techniques have been more praised than validated, and why rationalization has somehow been allowed to outrank validation in importance.

Needs and expectations

One reason for this situation is the acute need for measures of speaking ability that has been so often stated in the language testing literature, and pointed out as an obvious flaw by those from without. This has created an air of need and ready acceptance that makes it very hard to be objective when a measure appears that seems to fill the vacuum. We cannot ignore that the need for such measures, coupled with our lack of prowess in this area (which has been so often publicly admitted), has created a very favorable climate of validation for a measure such as the oral interview. We have very high expectations for it, and we think that it "works."

Again, we need to keep in mind that the point here is not whether we *think* it really works or not. Rather, the spirit of validation, honed on many past unfortunate experiences, demands that every test be assumed during developmental stages to be *not* valid until "proven" so to certain well-established levels and standards. These standards acknowledge the fact that once a test gets loose, once it is freed from the constraints of validation, it is almost impossible to catch again (cf. Buros, 1972: xxviii).

We should also keep in mind, however (especially with the current, popular tendency to view all testers as mercenaries or simple miscreants), that no language testing specialist has ever authored or supported a test that he or she knew was *not* valid. The severity of the spirit of validation is of course designed to prevent the misuse of measures, and recognizes that most measures are misused when they appear to satisfy an educational need, and when expectations for their validity are allowed to outrun actual validation. Much of the severity of the spirit of validation derives from this fact of measurement life: most test authors *are* trying to fulfill a felt need, and most test authors *are* sincerely convinced that their tests and testing theories *are* valid. The spirit of validation therefore not only tries to prevent the willful and careless misuse of tests; it also tries to protect the test constructor from his or her own self-confidence. Moreover, and perhaps most importantly, it tries to protect the test constructor and any future examinees from the too willing acceptance of a measure by those test users who impatiently argue that their practical needs are of primary importance, and that test validation, while nice, is not.

Face validity

One of the most striking aspects of the FSI Oral Interview and related approaches is that their primary appeal seems to derive from their high face validity. This attribute, however, is most often connected with "public relations" in the measurement literature. It is hardly one that a measurement specialist would offer to another as evidence that his measure does, in fact, reflect an extremely complex construct. The basic problem with the climate of validation for the oral

proficiency measures such as the oral interview is that they have already been assumed or declared to be "valid" upon grounds of face validity. Such "validation" does not meet generally accepted standards of validation.

Spolsky (1975b: 141) has pointed out that we in language testing are under no compulsion to accept standards in terms or tests from the educational and measurement literature. And at first glance, there does not seem to be any strong reason why we in language testing should not define terms to suit our problems and purposes as they arise. Nonetheless, both in principle and in practice, such terms and standards form a closely interrelated network of assumptions. Each necessarily affects others, and all are eventually realized in basic statistical procedures. Even a casual change in meaning can affect the entire network and cloud a clear view of validation (Peterson and Cartier, 1975).

The use and acceptance of the term "face validity" is therefore hardly a casual or unimportant matter. It forms the major claim for the validity of oral interview measures (e.g., Clark, 1978b: 225), yet is not recognized by the measurement tradition as having any bearing on a technical consideration of what a test measures. Rather, face validity can be considered to be appearance of validity in the eyes of the metrically-naive observer. Face validity "is not validity in the technical sense; it refers, not to what the test actually measures, but what it appears superficially to measure" (Anastasi, 1968: 104).

Although the casual phrases "test validation" and "validity of a test" are often used for convenience by language testing specialists, the metrically-naive observer is also not familiar with the fact that no test can be considered to be valid in itself. There are *degrees* of validity of measurement *procedures*. Each procedure can only be adjudged to have a certain degree of validity with respect to a specified purpose, examinee population, interpretation, and so on. No test possesses an inherent validity independent of these restrictions.

> The casual phrase *test validation* seems to imply that the score one interprets comes from a naked instrument. The instrument, however, is only one element in a procedure, and a validation study examines the procedure as a whole (Cronbach, 1971: 449).

The metrically-naive observer, then, tends to judge an oral interview by its appearance, as a yes-it-is or no-it-isn't question of validity, and tends to assume that a technique that is in any way similar to an FSI Oral Interview inherits this "inherent" worth, and does so independently of any change in purpose, etc.

We are all aware that the high face validity of the FSI Oral Interview has often been used as an accolade in discussions. There does not seem to be an equal awareness that by appealing to the naive observer's lack of testing sophistication, a climate negative to technical validation will result. We can neither claim, nor allow it to be assumed, that an interview measures "real-life speaking proficiency" until we have more evidence from construct validation studies to support this claim. Cronbach has stated that whenever "an educator asks, 'But what does the instrument really measure?' he is calling for information on construct validity" (1971: 463). The climate of validation for a convergent and dis-

criminant validation study is, of course, altered if the educator does *not* ask the question. Or if, when the question is raised, the technically untrained educator answers, "That's a silly question! Take a look at it yourself. It's obvious what it measures!"

Face validity is of course very good for public relations, but its seductive appeal is a danger to any objective examination of construct validity. Because face validity plays such a strong role in discussions of the validity of oral proficiency measures, Cronbach's (1970: 183f.) caveat is worth repeating:

> Adopting a test just because it appears reasonable is bad practice; many a "good-looking" test has failed as a predictor . . . such evidence as this (reinforced by the whole history of phrenology, graphology, and tests of witchcraft!) is strong warning against adopting a test solely because it is plausible. If one must choose between a test with "face validity" and no technicIly verified validity and one with technical validity and no appeal to the layman, he had better choose the latter.

Face and/or content validity

It is also important to note, however briefly, that the tendency in discussions of oral proficiency measures to collapse face validity and content validity into the same classification (mixing popular and technical) obscures and therefore confuses a very important distinction. This distinction is extremely important for convergent and discriminant validation studies. The canons for construct validation require that construct validity "must be investigated whenever no criterion or *universe of content is accepted as entirely adequate to define* the quality to be measured" (Cronbach and Meehl, 1972: 92; emphasis added). Therefore, it is very important that the distinction be offered here as it is assumed by convergent and discriminant approaches and as given in the *Standards*:

> To demonstrate the content validity of a set of test scores, one must show that the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definitions of the performance domain, the user's objectives, and the method of sampling are critical to claims of content validity (28).

> It should be clear that content validity is quite different from face validity. Content validity is determined by a set of operations, and one evaluates content validity by the thoroughness and care with which these operations have been conducted. In contrast, face validity is a judgment that the requirements of a test merely *appear* to be relevant (29).

Content validity and criterion-related validity

One view of content validity that constantly plagues language testers is that, although linguists are generally willing to admit that what constitutes "oral language proficiency" is far from being established, when it comes to measuring oral language proficiency, somehow this lack of an adequately defined universe of content is not as readily apparent (Stevenson, 1979). Similarly, the great problems attached to the search for a "more ultimate" criterion are generally under-

rated by those outside language testing. But to meet content validation standards, definitions of the performance domain must go far beyond the "you know what I mean" level. Also, the acceptance of a criterion involves the acceptance of that criterion as a better indicant of the performance domain than the measure in question.

In reference to oral languge proficiency measures, the acceptance of a universe of content as defining the variable is not at present possible. The best known statement of this problem remains Spolsky's (1968b) "What does it mean to know a language . . .?" The shortest and most direct summary of the argument is by Jakobovits:

> The question of what it is *to know a language* is not yet well understood and consequently the language proficiency tests now available and universally used are inadequate because they attempt to measure something that has not been well defined. (1970: 75).

The universe of content defining the construct "oral language proficiency" cannot yet be sufficiently described, and as a result the demands of content validation cannot be fulfilled.

The same arguments which have been used to point out the impossibility of specifying the "linguistic" elements of what it means to know a language can also be applied to our inability to specify what constitutes "real-life" sociolinguistic behavior. We can claim, but we cannot really demonstrate, that an oral interview constitutes "a representative sample of behaviors to be exhibited in the desired performance domain." As Fishman and Cooper (1978) point out, we in language testing have been very lax in even trying to specify those situations which we hope to predict (to). Jones (1978) has voiced similar criticism in connection with the oral interview.

Until we can more completely specify what constitutes the desired sociolinguistic behavior, we are still in the position of trying to sample something that has not been well defined. We can make lists of notional categories, for example, but we cannot know that any one is necessary for a certain situation, as we do not yet know how they interrelate or their respective weights in those interdependencies. In other words, whether what is to be sampled is seen as "linguistic" or "communicative," the same principle applies: we are still postulating interrelationships, we are still very much involved with theory, and therefore (sooner or later) are dealing with construct validity. The same arguments which have been used in connection with discrete-point tests of "language" proficiency can be applied to the attempt to have discrete-point tests of "real-life" communicative ability.

The lack of a criterion or set of criteria which could be accepted as entirely adequate is more often recognized as an obvious problem for the validation of oral proficiency measures. If a better criterion were available, why not use it at least for validation procedures, practical matters aside? The choice of a "more valid" criterion relies, in turn, upon a judgment of content validity, however, and eventually on the theory that underlies the selection of that construct.

Such problems in dealing with content and criterion-related validity are important, of course, as they point to the necessity of considering other validation approaches. They are equally important in the way that they affect the climate of validation which surrounds oral proficiency measures. The metrically-naive test user is much less likely to be aware of the problems connected with content and criterion-related validation than is the tester, and therefore more willing to use a measure that has been insufficiently validated. It should also be noted that the content of an oral proficiency test is less likely to be seen as self-explanatory by the linguist than it is by the nonlinguist. Together, these views can be said to work against the awareness that content and criterion-related validation are both insufficient for oral proficiency measures.

Direct and indirect measures

Another problem that directly affects the climate of validation surrounding oral proficiency measures such as the FSI Oral Interview, and which is interrelated with views of content and criterion-related validation, is the belief that oral interviews are "direct" measures, in that they somehow sample the performance domain directly and do not require criterion-related validation. As I have argued elsewhere (1975; 1977a), the by now familiar dichotomy between so-called "direct" and "indirect" measures (Clark, 1975) rests upon an assumption of what the "face valid" real-life situations represent *before* the behavior sample in effect becomes a measure.

The dichotomy neglects both the problem of sampling whatever it is that constitutes "oral language proficiency" and the fact that language cannot be assessed without the method of measurement leaving some traces, weak or strong, upon the language "trait". Carroll has stated (1968: 51) that "the single most important problem confronted by the language tester is that he cannot test competence in any direct sense; he can measure it only through manifestations of it in performance." There is no such thing as a direct test when only the language part of language testing is emphasized.

When the testing part of language testing is emphasized, there is even less reason to speak of a direct test of oral language proficiency. Even a "simple" observation, impressionistic or structured, involves a sampling of behavior, and the strong possibility that some effects of the method of observation will interfere with that observation. For example, after the "real-life language-use situations" are filtered through scoring/rating procedures they are not necessarily any closer to predicting real-life behavior than other types of tests where the testing effects are more obvious (often by intent).

To assume that an oral proficiency test such as an interview is somehow a direct test of oral proficiency is to ignore a very important point. This point, which is strongly stressed in convergent and discriminant validation theory, is that any test is a "trait-method unit."

> The assumption is generally made . . . that what the test measures is determined by the content of the items. Yet the final score . . . is a composite of effects resulting from the content of the item and effects resulting from the form of the item used (Cronbach, 1946: 475; quoted by Campbell and Fiske, 1959: 85).

Here again it is important to keep in mind that it is a measurement procedure that is validated, not just the "content" of a measure. The content must be drawn through test items, through a test form, approach, or technique, and of course through whatever scoring or rating procedures are used. As was stated earlier, we in language testing tend to favor the language part of language testing at the expense of the testing part. But when speaking of the "direct" nature of oral interviews, we should remember that both parts function in a measure, that they are interrelated and not easily separable. The possible effects of method and metrification on trait can neither be ignored nor underestimated.

Validity and utility

In the discussion so far I have attempted to show how various views of both popular and technical concepts can affect the climate of validation which surrounds oral proficiency measures. Much of the discussion has been concerned with pointing out how technical views of validity differ from popular views. There exists, however, a very basic difference in views of when it is necessary to validate a measure and, in fact, of what is meant by "validity." Unlike the other concepts, these basic differences seem to operate in the background of discussions about the validity of oral proficiency measures almost as unstated assumptions. Furthermore, the difference between the views of validity is not so much on the popular versus technical level; rather, the difference can be traced to various technical definitions of validity and, specifically, the degree to which the utility of a test can be considered to relate to its validity. This difference is a basic one and, in spite of its assumed rather than stated appearance in discussions, it is important because of the effects it can have on validation approaches and must be given more attention.

There exists in the literature of educational and psychological measurement a wide spread of views of what validity means in relation to a measure. One common view is that "in a very general sense, a measuring instrument is valid if it does what it is intended to do" (Nunnally, 1967: 75). A more extreme view of this same common definition is given by Edgerton (1949: 52; quoted in Ebel, 1961): "By 'validity' we refer to the extent to which the measuring device is useful for a given purpose." Such views can be contrasted to the one that states "a test is valid if it measures all of and only what the examiner wishes it to measure" (Anastasi, 1972: 77), and to Ingram's (1977: 18) "When a test measures that which it is supposed to measure, and nothing else, it is valid."

It is suggested that the first two definitions have been used, at least implicitly, in discussions of the FSI Oral Interview, and that the last two more

closely reflect assumptions which are required by a convergent and discriminant approach in that the emphasis is not only on what a test does measure, but on what it should not measure, as well. The importance of this distinction will be discussed in a following section which treats the Campbell and Fiske approach itself. At this point it is the entry of utility into the definition of validity (as in the Edgerton definition) that is of prime interest.

At the 1974 Washington Language Testing Symposium (Jones and Spolsky, 1975), for example, we can see the different views operating in discussions of the FSI Oral Interview. Wilds, for instance, states that

> the fact of the matter is that this system works. Those who are subject to it and who use the results find that the ratings are valid, dependable, and therefore extremely useful in making decisions about job assignments (35).

Later in the discussion session the second type of view of validity can be seen when Spolsky (40) asks whether or not there have been studies to show to what extent the interview does, in fact, predict performance in other kinds of real-life situations. Wild's answer (40f.) is that "this has not been systematically examined as far as I know." This situation for the most part still exists today. Jones, for example, while careful to point out the lack of validation studies (e.g., 1975: 4; 1977: 236), still would maintain that

> despite its acknowledged shortcomings, the oral interview remains the most useful and valid instrument for measuring spoken language proficiency (1978: 93).

It would appear then that some of the claim for the validity of the FSI-type measure rests upon an assumption that utility and practicality are, or should be, aspects of validity. It can be argued however that the question of utility is not, strictly speaking, related to validity, and that questions of utility must be kept separate. There are several reasons for this position.

First, if utility is allowed into definitions of validity, we are then in the position of questioning not only whether a measure is a measure of oral language proficiency, for example, but also of whether the *purpose* of the testing procedure conforms to the purpose of the test. It is, after all, only the "real" purpose of a testing procedure against which validity can be questioned. Here, it is perhaps useful to introduce the concept of "institutional validity," which can be seen as the extent to which a test fulfills a certain institutional need (e.g., selection, placement), irrespective of whether or not that test can be said to be a valid measure of some language ability. An example of a test with high institutional validity would be an admissions test which no longer is used to gather information about students' language abilities, but is used for gathering demographic information, or simply because it has become part of the ritual of admission procedures. Another example would be the inclusion of a certain type of test in a testing program simply because it has cosmetic or public relations functions. Finally, there is the all too common use of a test just because a test is required: the purpose of the testing is to test. A recent case I have heard of concerns a teacher who was told that regulations and tradition required that reading com-

prehension tests in a foreign language be in the form of oral examinations, only. In any case, he was "not to worry" as the examinations were largely a matter of form.

I have offered these examples not because they necessarily fit the situation of the FSI Oral Interview, but because when utility enters into definitions of validity, such questions must realistically be raised. We would be naive if we did not admit that many of those who use a measure use it because of administrative needs, rather than those more closely connected with language learning and teaching. Yet while this is often the case, a basic assumption of validation theory is that the stated purpose of a test is also, *in reality*, the purpose of the testing. When this is not the case, it makes more sense to speak of the misuse of a test (or testing) than of validity. In short, I would argue that the use and utility of a test must be dependent on the validity of that test (for a purpose that conforms to that test's purpose), rather than the other way around.

Another reason why we should be careful to keep practicality and utility separate from validity is that if they are allowed into the concept of validity, the concept becomes more associated with value than with what, after all, a measure is measuring. It should be noted that practicality when expressed as reliability does enter into concepts of validity to the extent that reliability and validity are interrelated. The relationship can be seen for example, in the fact that there are practical limits to the length of a test, and that the length of a test has a definite relationship to the degree of reliability that is possible. In turn, high reliability is a necessity for any degree of validity. This is expressed in the familiar measurement saw that a test can be reliable without being valid, but to be valid, it must be reliable.

The distinction between what is ideal and what is possible in real-life testing practice is nonetheless one of the basic differences that is present in concepts of validity and reliabilty. For example, if one were to obtain the test-retest reliability of the "best measure available," or the most practical test, this could not be interpreted as a "perfect" validity coefficient. The use of corrections for attenuation is also an example of this distinction between what is possible and what is ideal. Corrections for attenuation operate in the area of "*if* this measure *were* more reliable" (or even perfectly reliable), *then* this could be said about its validity, etc. It should be noted that the "if" cannot be discarded when the test and its validity are returned to the context of real-life testing practice, where the less than perfect reliabiltiy necessarily limits the validity of a measure.

The distinction between practicality and validity is also recognized in the basic logic of construct validation, which begins by assuming that no single measure is or can be a perfect reflection of a construct. Each of several measures may be an imperfect indicant of a construct or constructs (dependent of course on the stated purpose of a test), but no single measure can be a perfect indicant of any one, or all. Reliability problems aside, a test is a sample of some behavior, and our imperfect knowledge of that behavior dictates that our sample will

also be imperfect. Nonetheless, the ideal is to maximize the degree to which a measure reflects a construct.

Given the reasons I have sketched above, we must be extremely careful, if we are to speak of the construct validity of oral proficiency measures, to differentiate between practical decisions in test use and the claim that tests are valid because they are "useful" or because we do for practical reasons what must be done. Utility, as far as is possible, and reliability, as far as is possible, must be kept separate in discussions of validity. As has been shown, both practicality and reliability do impose limitations on validity in real-life testing contexts. It does not follow that the *concept* of validity, which while acknowledging the practical limitations assumes the ideal, should be constrained by this reality. To do so would be to limit our search for a more valid measure, and subsequently our understanding of the construct, oral language proficiency, itself.

Unitary versus divisible competence

A final area that must be considered because it so strongly affects the climate of validation for oral proficiency tests is the growing interest in the "unitary versus divisible competence" hypothesis. The basic question here is whether or not language proficiency is a unitary behavioral phenomenon, or whether, as "traditional" language testing models suggest, it is divisible (e.g., into the "four skills"). This dispute, which has emerged from basic questions of language testing theory (e.g., Carroll, 1961; Spencer and Holtzman, 1965; Spolsky, 1968b; Oller, 1973) and from questions dealing with the validation of the cloze (cf. Stevenson, 1978), has come to include oral proficiency measures such as the FSI Oral Interview "type" within its arguments and data bases.

There are several reasons why this fast-growing area of research can affect the planned validation study. First, much of this research is at least partially based on arguments which state that all language measures which purport to test language proficiency must be answerable to construct validation. Detailed support for this position can be found, for example, in Stevenson (1974), Petersen and Cartier (1975), or in Stevenson (1975). Their importance to the construct validation of oral proficiency measures is that they provide theoretical support for the position that content and criterion-related validities are insufficient.

Secondly, there is the fundamental question of whether the concepts of "speaking, listening, reading, writing" can be *measured* as separate or/but related constructs, and whether it is possible to *demonstrate* at the score level that they possess construct validity. Research involving these questions can be seen, for example, in Oller (1976a; 1976b), Oller and Hinofotis (1976), Oller and Perkins (in press), and Scholz et al. (1977). Critical examinations of the empirical research and associated theory are found, for example, in Upshur (1976), Sang and Vollmer (1978), and Vollmer (1978). Again, the importance for the planned

study is that arguments for the need to validate proficiency measures at the construct level of speaking, reading, etc., are presented in these efforts.

Thirdly, and I think most importantly, we must consider what effect on the planned validation study these several efforts might have. One possible effect which should not be underestimated is that because oral interviews have been used in several of these studies, and because their validity as measures of oral proficiency has often been assumed, any future attempts to validate measures such as the FSI Oral Interview will necessarily reflect upon these other studies as well. In other words, a wall of hypotheses has been built in some of these studies, and conclusions have been drawn as to the central hypothesis itself. Because oral interviews have been included, and because they have been assumed to be valid, they serve in a manner of speaking as building blocks in this wall of hypotheses. Any future questioning in theory or in data analyses of oral proficiency measures can therefore endanger this wall of hypotheses and, of course, any conclusions which have been reached.

Because of the fundamental nature of the unitary versus divisible hypothesis research, I assume that in any convergent and discriminant validation study involving oral proficiency measures there will be a strong tendency to state hypotheses and to interpret data with an eye to their significance to this previous research. This is especially likely because of the tendency at present in language testing to claim that valid language tests are those which conform to a certain linguistic theory (e.g., Oller, 1978: 52) or testing approach (e.g., discrete-point or integrative). Such definitions do not necessarily allow for the logic of validation that must be assumed for construct validation studies. This logic is that validity by decree is not possible; rather, the theory upon which a test is based is validated along with the test and *must* be subject to such validation. There is, then, the danger that assumptions of what a valid test should look like could be taken over into the planned study as facts, instead of remaining as basic questions. For example, to assume, because of theoretical arguments, that a cloze is a valid measure of "overall language proficiency" or that a "discrete-point" test battery is not, and on this basis to choose measures in a study, is to take as fact what should best remain as hypothesis. Furthermore, if we are to objectively consider the validity of oral proficiency measures, we must be aware that because such measures have been taken as valid in several unitary versus divisible hypothesis studies, this exerts some pressure on the spirit of validation.

Convergent and discriminant validation

Any understanding of convergent and discriminant validation first necessitates that certain views of validity, certain concepts, and, not least, certain attitudes on the part of the tester, be clarified. In the preceding sections of this paper an attempt has been made, by discussing the climate of validation that surrounds oral proficiency measures, to point out some of the more important

problems which must be clarified before convergent and discriminant validity is approached. The detailed arguments for the necessity of examining the construct validity of language proficiency measures can be found elsewhere, as was previously mentioned (e.g., Stevenson, 1974; Petersen and Cartier, 1975; Stevenson, 1975), and have of course been partly presented in previous sections of this paper. A brief summary of these arguments is nonetheless useful before turning to the convergent and discriminant approach itself.

The concept of construct validity, the necessity of validating the theory underlying a test (Noll and Scannell, 1972: 141), has been an important force in educational and psychological measurement theory since the early 1950s. The concept was originally derived from some of the problems related to the measurement of personality traits (Cronbach, 1971: 462), but its relevance to the broader area of educational measurement has become increasingly clear. As was noted previously, when an educator asks, "But what does the instrument really measure?", he is calling for information on construct validity (Cronbach, 1971: 463).

A problem that faced the early proponents of construct validity, and one that remains with us today, is that the concept represents in practice no single procedure. There is no "correlational coefficient of construct validity," for instance, corresponding to that which is so closely associated with the more familiar concept of criterion-related validation. Because constructs are being dealt with (constructs are in fact being "tested"), there is no clear-cut procedure which would yield a yes-it-does or no-it-doesn't answer. Rather, a logical orientation is involved, and it is basic to the concept that it *not* be identified with any single investigative procedure (Cronbach and Meehl, 1972: 92).

It is also basic to the concept that it not be seen simply as an alternative and separate approach to validation, but as one that is necessary when content or criterion-related approaches are either insufficient or inappropriate:

> When an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity because this is the only way to avoid the 'infinite frustration' of relating every criterion to some more ultimate standard . . . In content validation, acceptance of the universe of content as defining the variable to be measured is essential. Construct validation must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured (Cronbach and Meehl, 1972: 92).

As was argued earlier in the section on content and criterion-related validity, neither content nor criterion-related approaches can be seen to be sufficient for oral proficiency measures, and therefore a construct validation approach is called for.

Although the logic for construct validation would seem to be rather straightforward, those working with the planned study will no doubt find it necessary to defend the concept itself. This will probably be the case because, with a few exceptions, until the recent interest in unitary competence hypothesis studies little interest in construct validation has been apparent among language

testers. Until Heaton's (1975) text on language testing, no introductory-level textbook gave the concept any attention. Davies, one of the notable exceptions, stated as early as 1965 that "construct validity is what language tests need most of all . . ." (36), and mentioned it again in his introduction to the 1968 *Language testing symposium* volume (Davies, 1968). Valette (1968:114) gave the concept passing attention, but as far as I know, the only studies to give the concept specific attention prior to those dealing with the unitary competence hypothesis are those by Angoff and Sharon (1970) and Pike (1973). On the other hand, the questioning of the "four skills" classification has been widespread, and an implicit recognition of the concept of construct validity can be seen in a study by Spencer and Holtzman (1965). In spite of such efforts, it is apparent that much work still needs to be done in an effort to make the concept of construct validity better known in language testing.

Outside of language testing, the concept of construct validity has had much more influence, as has the related concept of convergent and discriminant validation. One reason that the concept of convergent and discriminant validity has been so influential is that in their 1959 paper, "Convergent and discriminant validation by the multitrait-multimethod matrix," Campbell and Fiske brought together into one conceptual framework concepts of validity that had existed only separately. According to Tapp and Barclay (1974: 440),

> the concept of convergent validity is identical to the rationale for traditional approaches to validation, especially criterion-related validity. That is, there should be substantial agreement between the test's measurement of its traits and the criterion measure of those traits.

As content validity is interrelated with an examination of the criterion's validity, content validation is also among the traditional approaches.

The novel and most important feature of the Campbell and Fiske approach is the emphasis on discriminant validity:

> Discriminant validity refers to the notion that *traits* should be distinguishable from each other when measured by different *methods*. The situation is evidenced when the agreement between different measurement procedures for a trait is greater than the intercorrelation between that trait and others within the same measurement procedure (Tapp and Barclay, loc. cit.).

It is the emphasis upon the effects of the interaction of both trait *and* method that is so important in the Campbell and Fiske approach. Similar trait measures can show substantial correlations not only because of similar traits, but *also* because of similar *methods* of measurement. Campbell and Fiske present their logic and trace their reasoning in four steps. Because these steps are logically ordered, and must be followed one after another, they are here presented together, and then commented on.

1. Validation is typically *convergent*, a confirmation by independent measurement procedures. Independence of methods is a common denominator among the major types of validity (excepting content validity) insofar as they are to be distinguished from reliability.

2. For the justification of novel trait measures, for the validation of test interpretation, or for the establishment of construct validity, discriminant validation as well as convergent validation is required. Tests can be invalidated by too high correlations with other tests from which they were intended to differ.

3. Each test or task employed for measurement purposes is a *trait-method unit*, a union of a particular trait content with measurement procedures not specific to that content. The systematic variance among test scores can be due to responses to the measurement procedures as well as responses to the trait content.

4. In order to examine discriminant validity, and in order to estimate the relative contributions of trait and method variance, *more than one trait* as well as *more than one method* must be employed in the validation process. In many instances it will be convenient to achieve this through a multitrait-multimethod matrix. Such a matrix presents all of the intercorrelations resulting when each of several traits is measured by each of several methods (1959: 81).

The comment that validation is typically convergent is more the case with language testing than with other areas where the trait does not have such a "self-evident" content. The typical approach has been to correlate the scores of one test with another language test, and if a positive relationship is found (often low or moderate) to assume that the first measure also measures what the second test does to the degree indicated by the correlation. For example, a reading comprehension test is correlated with another reading test serving as a criterion, and if a reasonably high relationship is found, it is assumed that the first test is also a "valid" test of reading comprehension.

That both could be indicants of another trait, or multiple traits, is often not considered. Somehow the familiar elements-by-skills matrix has been taken as a literal mapping of construct relationships, that is, it is assumed that a mutual exclusivity exists (instead of a representation of the assumption that certain interrelated traits can be separately *emphasized* in testing). Similarly, there is the strong tendency to assume that a test has some inherent one-to-one relationship with a trait. It is often taken for granted that a test can only be an indicant of one construct, and that *independent* of purpose and use, it still possesses that relationship. A test asking for definitions of words and labelled "vocabulary" could, of course, be used to judge a need for closure, or frustration, or "intelligence."

For much the same reasons, there is a lack of recognition that scoring (as it reflects purpose) determines what has been tested at the score level. The belated recognition (cf. Stevenson, 1977a) that "real-life" behavior must be scored by "real-life" criteria to measure "real-life" speaking proficiency is one indication of this problem. To use discrete-point concepts for scoring (e.g., accent, vocabulary, grammar, etc.) of integrative situations and then to claim that the scores represent "direct" behavior is an example of the overwhelming attention paid to "trait" as opposed to "method." The problem is simply that it has not yet been fully appreciated that trait and method *both* interact in determining what is being measured. What has often been assumed to be common (stable) trait variance could be common method variance (or parts of both).

Unfortunately very little attention in language testing has been paid to the last two logical steps in the Campbell and Fiske set. If oral language measures are to be judged to measure "oral language proficiency" and only "oral language proficiency," this must proceed from the full Campbell and Fiske approach. That is, validation must be based upon a demonstration of the convergence and discrimination of *traits*, irrespective of *methods*. The systematic variance contributed to a matrix by method must be accounted for or given neutral status. Method variance includes the effects of the form and format of the test, or in short, *whatever is not intended to be part of the construct definition*, yet is associated with the total measurement procedure. Since we can only "know" a construct such as oral language proficiency through some form of measurement (including our own observations), convergent *and* discriminant validation for both trait *and* method is mandatory.

It is therefore very important to realize that whether data is considered by examining tables of intercorrelations or through factoral analyses of them, conclusions cannot be reached about the construct validity of various measures included *unless* the possible complicating effects of method have been taken into account. Both method and trait can contribute to a correlational matrix without identifying themselves. To assume that only trait variance is at work is to introduce a built-in source of error into research designs and data interpretation. I would suggest, for example, that one very basic problem with most unitary competence hypothesis research to date is that the hypotheses stated (e.g., Oller, 1976a: 149 ff.) reflect only the first two steps in the Campbell and Fiske logic for convergent and discriminant validation. They seem to ignore the effects of method which must also be considered (cf. Stevenson, 1978). And yet as Campbell and Fiske have emphasized, it is far from atypical to find measures showing "an excessive amount of method variance" (1959: 94f.). In some cases this method variance even exceeds the trait variance. A "high" degree of convergence can be explained on grounds of common method variance as well as common trait variance. There is no way of judging unless different trait measures and different methods are paired and introduced for comparison and contrast.

Traits and methods

Unlike many other areas of testing, language testing has a very special and complex problem when it comes to traits and methods. This problem is simply that what is trait and what is method is very hard to distinguish, and what should be considered as trait and what should be considered as method is very hard to decide. In an interview, for example, a "normal" conversational question given by the examiner is both part of the trait and part of the testing method. It could be argued that such a question would be appropriate to a definition of the construct, oral language proficiency, and therefore relevant trait rather than irrele-

vant method. Other cases are much more problematic. At what point, for example, do the examiner's questions become more method than trait? When they cease to be likely to occur in "real-life" situations? Or should we assume that because few if any adult examinees are likely to forget for a moment that an interview is a test, and not a tête-à-tête, all questions within the interview are colored by "method"?

In other areas of proficiency testing, similar questions can also be raised. We are generally in agreement, for example, that oral/aural phenomena should not be a part of a reading comprehension test, as it is "reading" we wish to test and not speaking or listening abilities. At the same time, many of us while reading in a foreign language are aware that a "little voice" keeps us company in our heads as we read, mispronouncing each word out loud in our minds as we read "silently" on. Anyone who has administered reading tests is also familiar with the sound of a low hum, and the sight of lips moving. Similarly, there is the question in listening comprehension tests of how many examinees "write out" in their heads what they hear.

We will obviously need to spend much time in the study considering this problem. Perhaps the best strategy at this point is to proceed from the definition offered earlier, that method variance is whatever is not intended to be part of the construct definition. It is a complex problem, however, and is related to the familiar observation that linguistics is a field that must approach its subject matter through its subject matter. This is a major problem in language testing, and in language test validation as well. We are trying to measure something with tools that are made largely out of what we are trying to measure, and the problem is to separate the tool from the matter.

In whatever way this issue is resolved for the proposed study, it is clear that the basic interplay of method and trait will complicate our interpretation of test statistics. If, for example, a statistic used to estimate the reliability of a test assumes the discreteness of items, and the test is claimed to be valid partly because there *are* interdependencies among items, then the estimate given by the statistic will influence other conclusions as well by contributing to correlation coefficients. Or, for instance, if the assumption is made that oral interview judges are using the individual rating scales, and yet there is some indication that halo is operating (Stevenson, 1977a; Callaway, 1977; Mullen, 1978), there is a question of whether or not the interpretations based upon absolute definitions can be claimed to be valid, to the extent that the individual scale categories are reflected in the verbal descriptions.

Suggestions and directions

The development of the design for a convergent and discriminant matrix is beyond the scope of this paper; but as Campbell and Fiske stress, the logic dictates the steps rather than the other way around. It is the basic concepts

which are important, especially the discriminant validation requirements and, most importantly, the attention given to the effects of method variance. I am aware of only two studies which have specifically acknowledged the Campbell and Fiske approach and which have also taken it in its entirety, that is, paid specific attention to the effects of method. The first one, by Corrigan and Upshur (1978), is available only in a pre-publication copy (which is not for citation). Respecting this restriction, it can only be said here that this study does not concern itself specifically with oral proficiency measures, yet seems to strongly support the contention that method variance must be a consideration in any examination of convergent and discriminant validity. The second study, by Clifford (1978), is discussed by him elsewhere in this volume. Several other studies are underway, but the impact of convergent and discriminant studies upon language proficiency measures is still too limited to make any conclusions.

Judging from the few studies which have been attempted so far, it is fair to state that the problems encountered have been mainly related to meeting the basic requirements of the first steps in the Campbell and Fiske set. Corrigan and Upshur (1978), for instance, had to deal with rather low reliability estimates, and Clifford (1978) had to consider whether or not the independence of methods requirement was fulfilled. The question of how "high" the reliability of a test should be is of course a relative one. Nonetheless, unless a measure meets the generally accepted levels (by test type) — for example, in the 90's for a standardized test — it cannot be said to have met one of the most basic requirements for entry into a multitrait-multimethod matrix. To use the measure in such a matrix then would be to "skip" a required step. There is also no precise rule available that would tell us exactly when methods are independent. This is also a relative question, and one can only say that they should be as different as possible. The experience gained from studies such as those mentioned will of course make it much easier for those who follow to deal with similar problems.

In general, it has been observed that studies which use the Campbell and Fiske approach (outside of language testing) often do not find support for the construct validation of the measure in question. The most common reason is the failure to support the discrimination criteria. Whether or not this will prove to be the case within language testing is, of course, not possible to say. If subsequent convergent and discriminant studies do *not* tend to support the construct validity of oral proficiency measures, however, we can be assured that a great deal of attention will be given to examining their research designs. Each step must therefore be given full attention.

There have been many refinements made in the basic matrix since 1959, and these have not made the approach any less complex, or any less rigorous. A good example is a fairly recent paper by Tesser and Krauss (1976) that examines certain aspects of discriminant validation theory as it relates to the larger area of construct validation. In the course of their discussion they examine (and seemingly reject) "nonsolutions" such as various factoral analyses and corrections

for attenuation. At the same time, they point out some of the arguments concerning each.

This complexity of debate leads one to question if it might not be useful to state some "rules of the game" before the study begins. The simple statement, for instance, that the relation between two measures must be significant and be "sufficiently large to encourage further examination of validity" can become a problem, especially if there is an air of hoping that such will be the case. As has been shown, what constitutes independence of methods will certainly be a problem. One approach is to follow the general recommendation that the entire study be designed to directly counter the investigator's sympathies. The requirement that "several methods in one matrix should be completely independent of each other: there should be no prior reason to believe that they share method variance" (Campbell and Fiske, 1959: 103), might be approached, for example, by choosing those measures which the investigators feel should be *least* related by method, and then submitting them to outside judgment. In practice, of course, this requirement will remain something to reach for, rather than something that can be reached. But if this is done before the measures are given final approval for inclusion in the study, it would certainly increase the chances that the study will be independent of investigator bias.

For similar reasons, I would suggest that the study consider including outside psychometric specialists to interpret data, or suggest alternatives to data interpretation. It would also appear sensible to include in the matrix traits which would *not* be expected to be measured by the oral proficiency measures, yet which can also be matched with different distinct methods. Because we generally do expect a high level of correspondence among all language measures, care should be taken to have some measures which are as "mode" pure as is possible, so that upon later examination some conclusions might be offered as to trait and method effects. Care must also be taken to insure a wide range of candidates and educational backgrounds.

It would also be interesting to consider what can be called "nonreactive" measurement ("unobtrusive," etc.). There has been a too willing acceptance that such observational techniques are beyond the language testing pale. No one would deny that there are great difficulties involved: the standardization of observational techniques, the large numbers involved, the necessity of establishing a large common set of nonreactive measures for the examinees, and, foremost, the need to have the examinee approve of such research beforehand and still do so without losing the "unobtrusiveness." Nonetheless, the inclusion of several such measures would greatly enhance the entire matrix, since how individuals speak when they are not being "tested" is a very important aspect of any attempt to validate oral proficiency measures.

As I have tried to point out throughout this paper, there is a wide range of terminological usage, definitions, and not a little jargon that confuses the entire area of oral proficiency measures. Some agreement must be reached, because

the arguments that have been given for the necessity of examining the convergent and discriminant validation of oral proficiency measures follow from such terminological distinctions and related theories. It is interesting to note in this respect that twenty years after the Campbell and Fiske approach appeared, we are beginning to appreciate its importance. Its basic logical approaches have not changed as much as our theories. Validation is a logical process, and if the "ifs" and "therefores" are casually discarded, or disregarded to suit a theory, so is the process. Moreover, one cannot argue from within the tradition when it is convenient, and from without when it is not. A consistently observed set of terminological usage, etc., is a basic requirement for the planned study.

Conclusion

In this paper I have emphasized some of the basic problems which must be faced in approaching a convergent and discriminant validation study of oral proficiency measures. There is a great tendency to rush to measurement, to gather data first, and to state hypotheses and definitions later. The fact that this colloquium was called together is in itself a good sign that this tendency will be resisted. On the other hand, and as I have intentionally stressed, we must guard against a positive bias in judging the validity of oral proficiency measures, especially the oral interview. To adapt a comment made about another test at the 1974 Washington Symposium, we would all dearly love to have what we've been pretending to have: an oral proficiency test validated in some legitimate fashion. The belief that we can counter our own desire to prove our own theories is, even from the viewpoint of recent testing research, highly doubtful. Faced with this situation, Levine (1974) and other testing historians have suggested that adversarial procedures might well be applied to the validation of tests. We should seriously consider adopting such procedures. A formally specified adversarial component using outside critics would greatly strengthen the planned study, and give more credibility to whatever conclusions are reached.

Finally, there is an impatience with theory whenever someone finds something that works, and someone else says "prove it!" However, when we decide to adopt a validation approach such as the convergent and discriminant one, we are saying that although we have a test that works within practical or reasonable limits, we are not content with "practice." We are saying that by examining the gap between best practice and best theory, and that by demanding a better measurement for the width of the gap, we might learn something about the construct in the process. The convergent and discriminant approach is one of the most severe in its demands. Few tests that have entered a multitrait-multimethod matrix have come out as good-looking as when they went in. Nonetheless, as I have argued in this paper, if we wish to go beyond faith in our measures, and beyond their face validity, we must also be willing to take a very critical look at where we stand, and how far validation must go.

REFERENCES

American Psychological Association. 1974. *Standards for educational & psychological tests*. Washington, D.C.: APA.

Anastasi, A. 1968. *Psychological testing*, 3rd ed. New York: Macmillan.

_____. 1972. Some current developments in the measurement and interpretation of test valiuation. In V. H. Noll et al., eds., *Introductory readings in educational measurement*. Boston: Houghton Mifflin. 77-89.

Angoff, W. H. and A. T. Sharon. 1970. *A comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants to U.S. colleges*. ETS Research Bulletin, RB-70-8. Princeton: Educational Testing Service.

Buros, O. K., ed. 1972. *The seventh mental measurements yearbook*. Highland Park, N.J.: Gryphon Press.

Callaway, D. R. 1977. Accent and the assessment of ESL proficiency. In J. E. Redden, ed., *Proceedings of the First International Conference on Frontiers in Language Proficiency and Dominance Testing, held at Southern Illinois University at Carbondale, April 21-23, 1977*. (Occasional Papers on Linguistics, no. 1.) Carbondale, Ill.: Dept. of Linguistics, Southern Illinois University. 163-177.

Campbell, D. T. and D. W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin 56*, 2: 81-105.

Carroll, J. B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen and R. N. Campbell, eds., *Teaching English as a second language*, 2nd ed. 1972. New York: McGraw-Hill. 313-321.

_____. 1968. The psychology of language testing. In A. Davies, ed., *Language testing symposium: a psycholinguistic approach*. London: Oxford University Press. 46-49.

Clark, J. L. D. 1972. *Foreign language testing: theory and practice*. Philadelphia: The Center for Curriculum Development.

_____. 1975. Theoretical and technical considerations in oral proficiency testing. In R. L. Jones and B. Spolsky, eds., 1975: 10-24.

_____, ed. 1978a. *Direct testing of speaking proficiency: theory and application*. Princeton: Educational Testing Service.

_____. 1978b. Interview testing research at Educational Testing Service. In J. L. D. Clark, ed., 1978a: 211-228.

Clifford, R. T. 1978. Reliability and validity of language aspects contributing to oral proficiency of prospective teachers of German. In J. L. D. Clark, ed., 1978a: 191-210.

Corrigan, A. and J. A. Upshur. 1978. Test method and linguistic factors in foreign language tests. Paper presented at the 1978 TESOL Convention, Mexico City.

Cronbach, L. J. 1946. Response sets and test validity. *Educational and Psychological Measurement 6*: 475-494.

———. 1970. *Essentials of psychological testing*, 3rd ed. New York: Harper & Row.

———. 1971. Test validation. In R. L. Thorndike, ed. *Educational measurement*, 2nd ed. Washington, D.C.: American Council on Education. 443-507.

Cronbach, L. J. and P. E. Meehl. 1972. Construct validity in psychological tests. In V. H. Noll et al., eds. *Introductory readings in educational measurement*. Boston: Houghton Mifflin. 90-121.

Davies, A., ed. 1965. Language proficiency testing. *Report on sixth meeting of International Conference on Second Language Problems*. London: English Teaching Information Centre. 33-42.

———, ed. *Language testing symposium: a psycholinguistic approach*. London: Oxford University Press.

Ebel, R. L. 1961. Must all tests be valid? *American Psychologist 16*: 640-647.

Fishman, J. A. and R. L. Cooper. 1978. The sociolinguistic foundations of language testing. In B. Spolsky, ed., 1978: 31-38.

Heaton, J. B. 1975. *Writing English language tests: a practical guide for teachers of English as a second language*. London: Longman.

Ingram, E. 1977. Basic concepts in testing. In J. P. B. Allen and A. Davies, eds. *Testing and experimental methods. The Edinburgh course in applied linguistics*, Vol. 4. London: Oxford University Press. 11-37.

Jakobovits, L. A. 1970. *Foreign language learning: a psycholinguistic analysis of the issues*. Rowley, Mass.: Newbury House.

Jones, R. L. 1975. Testing language proficiency in the United States government. In R. L. Jones and B. Spolsky, eds., 1975: 1-9.

———. 1977. Testing: a vital connection. In J. K. Phillips. ed. *The language connection: from the classroom to the world*. Skokie, Ill.: National Textbook Co. 237-265.

———. 1978. Interview techniques and scoring criteria at the higher proficiency levels. In J. L. D. Clark, ed., 1978a: 89-102.

Jones, R. L. and B. Spolsky, eds. 1975. *Testing language proficiency*. Arlington: Center for Applied Linguistics.

Kerlinger, F. N. 1964. *Foundations of behavioral research: educational and psychological inquiry*. New York: Holt, Rinehart and Winston.

Levine, M. 1974. Scientific method and the adversary model: some preliminary thoughts. *American Psychologist 29*: 661-677.

Mullen, K. A. 1978. Determining the effect of uncontrolled sources of error in a direct test of oral proficiency and the capability of the procedure to detect improvement following classroom instruction. In J. L. D. Clark, ed., 1978a: 171-189.

Noll, V. H. and D. P. Scannel. 1972. *Introduction to educational measurement*, 3rd ed. Boston: Houghton Mifflin.

Nunnally, J. C. 1967. *Psychometric theory*. New York: McGraw-Hill.

Oller, J. W. 1973. Pragmatic language testing. *Language Sciences 12*: 7-12.

_____. 1976a. A program for language testing research. In H. D. Brown, ed. *Papers in second language acquisition. Language Learning*, Special Issue No. 4: 141-166.

_____. 1976b. Language testing. In R. Wardhaugh and H. D. Brown, eds. *A survey of applied linguistics*. Ann Arbor: The University of Michigan Press. 275-300.

_____. 1978. Pragmatics and language testing. In B. Spolsky, ed., 1978: 39-57.

Oller, J. W. and F. B. Hinofotis. 1976. Two mutually exclusive hypotheses about second language ability: factor analytic studies of a variety of language tests. Unpublished paper delivered at the winter meeting of the Linguistic Society of America.

Oller, J. W. and K. Perkins. In press. *Language in education: testing the tests*. Rowley, Mass.: Newbury House.

Perren, G. 1968. Testing spoken language: some unsolved problems. In A. Davies, ed., 1965: 107-116.

Petersen, C. R. and F. A. Cartier. 1975. Some theoretical problems and practical solutions in proficiency test validity. In R. L. Jones and B. Spolsky, eds., 1975: 105-118.

Pike, L. W. 1973. An evaluation of present and alternative item formats for use in the Test of English as a Foreign Language. (Draft) Princeton: Educational Testing Service.

Sang, F. and H. J. Vollmer. 1978. *Allgemeine Sprachfähigkeit und Fremdsprachenerwerb. Zur Struktur von Leistungsdimensionen und linguistischer Kompetenz des Fremdsprachenlerners*. Berlin: Max-Planck-Institut für Bildungsforschung.

Scholz, G. et al. 1977. Is language ability divisible or unitary?: a factor analysis of 22 English proficiency tests. Paper presented at the 1977 TESOL Convention, Miami.

Spencer, R. E. and P. D. Holtzman. 1965. It's composition — but is it reliable? *College Composition and Communication*, May: 117-121.

Spolsky, B. 1968a. Language testing: the problem of validation. *TESOL Quarterly* 2, 2: 88-94.

_____. 1968b. What does it mean to know a language, or how do you get someone to perform his competence? Paper presented at the Second Conference on Foreign Language Testing at the University of Southern California.

_____. 1975a. Language testing: art or science? Paper presented at the Fourth AILA Congress, Stuttgart.

_____. 1975b. Concluding statement. In R. L. Jones and B. Spolsky, eds., 1965: 139-143.

———, ed. 1978. *Approaches to language testing*. Advances in Language Testing Series: 2. Arlington, Va.: Center for Applied Linguistics.

Stevenson, D. K. 1974. A preliminary investigation of construct validity and the Test of English as a Foreign Language. Ph.D. dissertation. Albuquerque, University of New Mexico.

———. 1975. Construct validation and language proficiency measurement. Paper presented at the Fourth AILA Congress, Stuttgart. (To appear in *Language testing*, AILA.)

———. 1977a. Problems of foreign accent and native speaker bias in language proficiency measurement. Paper presented at the AILA/TESOL Meeting on Language Testing in Miami.

———. 1977b. Language testing and academic accountability: redefining the role of language testing in language teaching. Paper presented at the Eighth GAL Congress, Mainz. (To appear in IRAL.)

———. 1978. Face validity and loss of faith: some effects of recent cloze research on traditional views of language proficiency. Paper presented at the Ninth GAL Congress, Mainz. (To appear in *Kongreßberichte der 9. Jahrestagung der GAL.*)

———. 1979. Problems and practice in language testing: the view from the university. Paper presented at the 1979 German International Symposium on Language Testing. (To appear in *Lingua et Signa*, Vol. 1. Bern: Peter Lang Verlag.)

Tapp, G. S. and J. R. Barclay. 1974. Convergent and discriminant validity of the Barclay Classroom Climate Inventory. *Educational and Psychological Measurement* 34, 2: 439-447.

Tesser, A. and H. Krauss. 1976. On validating a relationship between constructs. *Educational and Psychological Measurement* 36: 111-121.

Upshur, J. A. 1976. Discussion of Oller's "A program for language testing research." In H. D. Brown, ed. *Papers in second language acquisition. Language Learning*, Special Issue No. 4: 167-174.

Valette, R. M. 1968. Evaluating oral and written communication: suggestions for an integrated testing program. *Language Learning*, Special Issue No. 3: 111-124.

Vollmer, H. J. 1978. Evidenz für einen allgemeinen Sprachfähigkeitsfaktor? Paper presented at the Ninth GAL Congress, Mainz.

Wilds, C. P. 1975. The oral interview test. In R. L. Jones and B. Spolsky, eds. 1975: 29-44.

Convergent and Discriminant Validation of Integrated and Unitary Language Skills: The Need for a Research Model

Ray T. Clifford
CIA Language School

Abstract. Correlational studies establishing the validity of language skill tests have traditionally described how well the tests converged, i.e., yielded equivalent results. Convergent and discriminant validation is a logical extension of these traditional procedures; but since it requires evidence of discriminant as well as convergent validation it is ideal for the more rigorous, and functionally more important, problem of establishing the construct validity of language skill tests. A re-examination of examples drawn from the literature shows that studies claiming convergent test validity consistently fail to demonstrate evidence of discriminant or construct validity for the traits they purport to measure. These failures may be the result of error variance introduced by testing and rating methods and/or by attempts to measure skills which are based on shared rather than unique contributing elements. Suggestions are given for minimizing both method and specification error variance in convergent and discriminant language validation studies.

A recent "state of the art" article by Cooley (1978) reminds educational researchers of the need for explanatory models in observational studies, and once again impressively demonstrates that "a correlation does not an explanation make." Convergent and discriminant validation as outlined by Campbell and Fiske (1967) is indeed a correlational procedure, but it is also noteworthy in that it presupposes an underlying explanatory research model.

The first part of convergent and discriminant validation is a generally accepted validation procedure. Cronbach (1971) describes convergent validation when he suggests that test validity can be estimated by computing the correlation between that test and another independently developed test of the same trait. As the second part of the name implies, convergent and discriminant validation merely adds an additional requirement that one be able to discriminate

among the correlations generated by different methods of measuring the same and different traits.

Thus the procedure presupposes a model hypothesizing the existence of more than one trait to be measured and more than one method of measuring those traits. It then requires (1) that separate methods of measuring the same trait correlate more highly with one another than they do with other traits measured by *different* methods and (2) that ideally, separate measures of the same trait correlate more highly with one another than with different traits measured by the *same* method. The benefits of adding this second requirement can best be demonstrated with actual examples. In the area of language proficiency the general term "trait" has often been equated with the four skills of listening, speaking, reading, and writing. As part of a study to validate the *MLA Cooperative Foreign Language Tests*, Myers and Melton (1964) compared faculty ratings of NDEA Workshop participants with the participants' MLA test scores in these four skill areas. Correlations excerpted from that study are reproduced in Table 1 to provide an example of the *multitrait-multimethod* correlation matrix needed to illustrate the points made by Campbell and Fiske.

TABLE 1
A Multitrait-Multimethod Matrix of Correlations from the Study by Myers and Melton

German Institutes (N=312)

		MLA Proficiency Tests				Faculty Ratings			
		Listening	Speaking	Reading	Writing	Listening	Speaking	Reading	Writing
MLA Proficiency Tests	Listening					.69	.74	.69	.72
	Speaking	.83				.66	.74	.66	.71
	Reading	.86	.82			.67	.69	.67	.69
	Writing	.79	.85	.86		.62	.68	.66	.72
Faculty Ratings	Listening	.87							
	Speaking	.86	.86						
	Reading	.82	.85	.84					
	Writing								

MLA Proficiency Tests — Faculty Ratings

The ratings by NDEA faculty members and scores on MLA proficiency tests represent two different methods of measuring language proficiency in each of the four skills or traits to be validated. As mentioned, *convergent validation* requires high positive correlations between the two separate measures of the same traits. These validity correlations are bold-faced in Table 1.

Although not perfect, these correlations are substantial and, reported in isolation, they were interpreted as evidence of concurrent validity for the two measuring procedures.

The additional requirement of discriminant validation forces additional comparisons which allow more accurate interpretation of these correlations in terms of *construct validity*. A comparison of the correlations in the bold-faced *validity diagonal* in Table 1 with correlations in the two adjacent *heterotrait-heteromethod* triangles (enclosed in dashed lines), shows that none of those validity coefficients consistently exceed correlations of that skill with other theoretically distinct proficiency skills. In addition, all of the validity correlations fall far short of matching the correlations in the *heterotrait-monomethod* triangles (enclosed in solid lines). Thus the data from this study fail to give evidence of construct validity for the concept of distinct listening, speaking, reading, and writing skills.

One advantage of using a convergent and discriminant validation procedure is that the matrix also gives some clues as to the sources of error variance present in the assessment procedures used. In the Myers and Melton study the comparatively high correlations in each of the *monomethod triangles* indicate the likelihood that shared method variance and not merely trait similarities contributed to individual scores and ratings. The high correlations in the off-diagonal *heteromethod triangles* might be the result of several factors, such as trait instability and lack of reliability in testing and scoring procedures. The presence of one or more of those factors could mean the methods used may not be adequate for measuring the trait. On the other hand it could also be that the second criterion of convergent and discriminant validation was not met because of specification error in the research model. That is, in the words of Campbell and Fiske (1967: 300), the trait measured "is not a functional unity."

The concept of "functional unity" is critical in testing language skills. It deserves special attention in a convergent and discriminant study because it may be that some aspects of language proficiency are shared across the language skills of listening, speaking, reading, and writing. Stevenson (1974) found evidence of this when he applied the principles of convergent and discriminant validation to three methods of testing students' proficiency in English as a second language. He tested 46 foreign students with an oral cloze test of listening comprehension; a noise test of listening comprehension; and the *Test of English as a Foreign Language*, which includes subtests of listening comprehension, English structure, vocabulary, reading comprehension, and writing ability.

The results showed that the oral cloze scores correlated much higher with

scores on English structure, reading comprehension, and writing ability than with the TOEFL listening comprehension score. This unexpected result raised the question of what skills are tested by an oral cloze test? To answer this question, a factor analysis with varimax rotation was performed and two factors were identified. The first factor correlated highly with all of the tests used, but the second factor correlated highly with only the three listening comprehension tests. Stevenson (1974: 126) concluded that "factor B is tentatively identified as a listening factor and factor A as the familiar general language proficiency."

The "familiar general language proficiency" to which Stevenson refers is labeled by Carroll (1973: 11-12) as one of the "persistent problems" of foreign language testing and as a "paradox that the more we attempt to measure *different* language skills, . . . the higher the correlations among the skills [become]." These high correlations have led some to the conclusion that there is in fact a single general foreign language skill. Oller (1976) has used factor analysis procedures, for instance, to develop substantial evidence indicating the existence of a general proficiency factor. Carroll (1974), however, gives three reasons why this conclusion can be questioned. In the first place, high intercorrelations among language skills are generally found only where adequate instruction has been given in all those skills. Secondly, high correlations do not preclude significant difference in *relative levels* of proficiency in each of the skill areas. His third reason is of greatest import for language proficiency studies: The language skills being tested are what Carroll (1973: 12) calls "integrated" language skills, which all "depend (or should depend) on a wide variety of detailed competencies in particular aspects of the language — its phonology, spelling, grammar, lexicon, and so forth."

Carroll's third reason is supported by the form of the generally accepted languages testing model proposed by Lado (1961), Cooper (1965), Carroll (1968), Harris (1969) and Valette (1971), and which is found in the rating criteria of the MLA proficiency tests, the FSI interview, and many other oral tests. This general model can be represented schematically by a test blueprint matrix as in Figure 1.

Although there is general agreement on a language testing model consisting of four basic language skills and contributing language aspects in each of those skills, several variations on this model have been proposed. Carroll (1968: 57) suggests the need to measure integrated language performance and proposes that for speaking tests this integrated performance is observable as "oral speaking fluency." Cooper (1965: 336-37) adds a third dimension to the basic model to test different levels of language usage such as "formal" and "informal." Valette (1971) submits a model which includes both developmental and communication objectives. However, because "communication" is listed on the vertical axis of her model, rather than as a culminating objective following the developmental objectives on the horizontal axis, many empty and improbable cells are generated. Despite differing details in the specific models proposed, these researchers

FIGURE I
Language Testing Model

LANGUAGE ASPECT / SKILL	Pronunciation or Orthography	Grammar	Vocabulary	Other
Listening				
Speaking				
Reading				
Writing				

agree that there are aspects of language proficiency which may be shared across language skills. These models all support Carroll's argument that the language skills of listening, speaking, reading, and writing are not "functional unities" and provide a compelling explanation for the high intercorrelations found in many studies among skill scores on language tests. It might therefore be expected that tests of different language skills, which, however, test the same underlying language aspect or aspects, could yield comparable results. Such an interpretation is certainly compatible with the findings of Carroll's study (1967) which compared FSI proficiency ratings with MLA proficiency test scores. In German, for instance, the scores of the 39 teachers tested yielded a correlation of .82 between the MLA speaking test scores and the FSI rating of speaking proficiency. Although a correlation of .82 is substantial, drawing a conclusion about the validity of these measurement procedures from that statistic is complicated by the fact that the FSI *oral* interview ratings correlated equally highly (.86) with the MLA *writing* test scores. Carroll (1967: 13) calls this circumstance "unfortunate," but offers no explanation. A plausible explanation may be found in the scoring and rating systems used in each of the testing procedures. In direct contrast to the MLA speaking test, pronunciation in the FSI rating procedure is only minimally considered in determining the interviewee's proficiency ratings (Wilds, 1975: 32); and both the FSI oral interview and the MLA *writing* test scores are largely based on grammatical accuracy.

Of course, this explanation hinges on the existence of hypothesized contributing language aspects in each of the language skills. Some support for this theory was found in a study done by the author in German (Clifford, 1978). Two measures of oral proficiency, the speaking portion of the MLA Cooperative Foreign Language Test and an oral interview for measuring Teacher Oral Proficiency (TOP), were used to measure four language aspects thought to contribute to oral proficiency: grammar, vocabulary, pronunciation, and fluency.

Despite high test reliabilities evidence was found for convergent but not for discriminant validation of the contributing language aspects. As can be seen from the multitrait, multimethod correlation matrix in Table 2, the language aspect correlations in the validity diagonal do not consistently exceed the correlations of that language aspect with other aspects measured by the same method. This indicates that the testing procedures themselves introduced method-specific error variance into the rating procedures.

TABLE 2
Multitrait, Multimethod Convergent and Discriminant Validation Matrix
(N = 47 for all variables)

Test	Language Aspect								
MLA	Grammar	—				Correlations in the validity diagonal			
MLA	Vocabulary	.876	—			are bold-faced			
MLA	Pronunciation	.882	.775	—		All correlations in this matrix are			
MLA	Fluency	.845	.946	.731	—	significant at the $p<.001$ level			
TOP	Grammar	**.810**	.827	.752	.783	—			
TOP	Vocabulary	.744	**.816**	.683	.796	.876	—		
TOP	Pronunciation	.741	.670	**.788**	.643	.838	.740	—	
TOP	Fluency	.687	.802	.657	**.819**	.864	.825	.731	—
		MLA Gr.	MLA Vo.	MLA Pr.	MLA Fl.	TOP Gr.	TOP Vo.	TOP Pr.	TOP Fl.

In an effort to control for this unwanted method variance, as well as for trait instability and inter-rater error variance, other data from the same study were also analyzed. The mean scores assigned students on independent first and second ratings of the same test administration were correlated and the multitrait, multirating matrices shown in Tables 3 and 4 were created and inspected following the pattern of convergent and discriminant validation. Only under these controlled conditions, with high intra-rater reliability of mean scores on each of the language aspects, were the criteria met for distinguishing grammar, vocabulary, pronunciation, and fluency as identifiable aspects of oral proficiency. Table 3 reveals no exceptions to the ideal requirements of convergent and discriminant validation of the four language aspects using mean scores on the TOP interview. Similarly, the correlated mean scores from the MLA speaking test in Table 4 show only one minor flaw: the correlation of the second rating of vocabulary with the second rating of grammar exceeds the correlation between first and second rating of grammar by .001. Thus there is some evidence for the existence and measurability of contributing language aspects *within* a given testing method. These differences, however, were not demonstratable across testing procedures, being evidently obscured by the introduction of method variance into the validation matrix.

TABLE 3
TOP Interview Multitrait, Multirating Convergent and Discriminant Validation matrix

Test Rating	Language Aspect								
First	Grammar	—				Correlations in the validity diagonal are bold-faced			
First	Vocabulary	.876	—						
First	Pronunciation	.838	.740	—		All correlation in this matrix are significant at the p<.001 level			
First	Fluency	.864	.825	.731	—				
Second	Grammar	**.939**	.832	.824	.829	—			
Second	Vocabulary	.883	**.943**	.799	.855	.891	—		
Second	Pronunciation	.829	.750	**.909**	.722	.810	.805	—	
Second	Fluency	.814	.716	.694	**.908**	.813	.791	.722	—
		1st Gr.	1st Vo.	1st Pr.	1st Fl.	2nd Gr.	2nd Vo.	2nd Pr.	2nd Fl.

TABLE 4
MLA Speaking Test Multitrait, Multirating Convergent and Discriminant Validation Matrix

Test Rating	Language Aspect								
First	Grammar	—				Correlations in the validity diagonal are bold-faced			
First	Vocabulary	.876	—						
First	Pronunciation	.882	.775	—		All correlations in this matrix are significant at the p<.001 level			
First	Fluency	.845	.946	.731	—				
Second	Grammar	**.937**	.901	.837	.890	—			
Second	Vocabulary	.856	**.953**	.769	.915	.938	—		
Second	Pronunciation	.853	.758	**.942**	.743	.869	.802	—	
Second	Fluency	.795	.914	.707	**.963**	.886	.926	.739	—
		1st Gr.	1st Vo.	1st Pr.	1st Fl.	2nd Gr.	2nd Vo.	2nd Pr.	2nd Fl.

Implications and recommendations

A convergent/discriminant validation procedure is especially suited for establishing construct validity of hypothesized traits. However, it must be remembered that the procedure is a very demanding one in that it requires a multitrait and multimethod approach. If assessment procedures are not precise they will introduce their own "methods" error variance. If the trait to be validated is not in fact a "functional unity," but shares aspects with other measured traits, "specification" error variance will cause inter-trait correlations to be spuriously high. Existing language proficiency studies provide ample evidence that, to be successful, a language skill validation study must use reliable assessment procedures and must be based on a language testing model which identifies those

aspects of language proficiency which overlap language skill areas. To ignore these requirements only serves to increase the error variance in the study and reduce the likelihood of covergent and discriminant validation. It is therefore recommended that:

1. An explanatory research model be developed which specifies the functional unity or language aspect to be validated.
2. If the possibility exists that the language aspect to be measured is not specific to any one language skill, instrumentation be designed to measure that aspect within and across language skills.
3. Efforts be made to minimize error variance in test scores which can result from extraneous factors such as "halo" effects in rating, trait instability over time, and general lack of reliability in testing and scoring procedures.

REFERENCES

Campbell, Donald T. and Donald W. Fiske. 1967. Convergent and discriminant validation by the multitrait-multimethod matrix. In William A. Mehrens and Robert L. Ebel, eds. *Principles of educational and psychological measurement*. Chicago: Rand McNally. 273-302.

Carroll, John B. 1967. *The foreign language attainments of language majors in the senior year: a survey conducted in U.S. colleges and universities*. Cambridge, Mass.: Graduate School of Education, Harvard University. EDRS: ED 013 343.

_____. 1968. The psychology of language testing. In Alan Davies, ed. *Language testing symposium: a psycholinguistic approach*. London: Oxford University Press. 46-49.

_____. 1973. Foreign language testing: Will the persistent problems persist? In Maureen Concannon O'Brien, ed. *ATESOL testing in second language teaching: new dimensions*. Dublin, Ireland: The Dublin University Press. 6-17.

Clifford, Ray T. 1978. Reliability and validity of language aspects contributing to oral proficiency of prospective teachers of German. In John L. D. Clark, ed. *Direct testing of speaking proficiency: theory and application*. Princeton, N.J.: Educational Testing Service. 193-209.

Cooley, William W. 1978. Explanatory observational studies. *Educational Researcher 7*, 9: 9-15.

Cooper, Robert L. 1972. Testing. In Harold B. Allen and Russell N. Campbell, eds. *Teaching English as a second language: a book of readings*, 2nd ed. New York: McGraw-Hill. 330-46.

Cronbach, Lee J. 1971. Test validation. In Robert L. Thorndike, ed., *Educational measurement*, 2nd ed. Washington, D. C.: American Council on Education. 443-507.

Harris, David P. 1969. *Testing English as a second language*. New York: McGraw-Hill.

Lado, Robert. 1961. *Language testing: the construction and use of foreign language tests*. London: Longmans, Green. (Reprinted 1965. New York: McGraw-Hill.)

Myers, Charles T. and Richard S. Melton. 1964. *A study of the relationship between scores on the MLA Foreign Language Proficiency Tests for Teachers and Advanced Students and ratings of teacher competence*. Princeton, N.J.: Educational Testing Service. EDRS: ED 011 750.

Oller, John W. 1976. Evidence for a general language proficiency factor: an expectancy grammar. *Die Neueren Sprachen 2:* 165-174.

Stevenson, Douglas K. 1974. A preliminary investigation of construct validity and the Test of English as a Foreign Language. Ph.D. dissertation. The University of New Mexico. DAI 36 (1975), 3: 1352-A.

Valette, Rebecca M. 1971. Evaluation of learning in a second language. In Benjamin S. Bloom, J. Thomas Hastings, and George F. Madaus, eds., *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill. 815-53.

Wilds, Claudia P. 1975. The oral interview test. In Randall L. Jones and Bernard Spolsky, eds. *Testing language proficiency*. Arlington, Va.: Center for Applied Linguistics. 29-44.

Structure of the Oral Interview and Content Validity

Pardee Lowe, Jr.
CIA Language School

Abstract. This paper suggests that use by interviewers of a deliberate, prearranged, and consistent overall structure, comprising Warm-Up, Level Check, Probes, and Wind-Up, can strengthen the content validity of interview tests. Moreover, the flexibility necessary for elicitation is increased if an established battery of well-structured tasks exists for candidates to perform. However, consistent use of the same topic in several interviews could lead to test compromise. Therefore, the recommendation is repeated use of underlying types of tasks and questions, but with different topics in different interviews, thus maintaining content validity and flexibility but avoiding test compromise. The paper also suggests that certain question types are useful at specific levels and presents samples for Levels 0+, 1, 2, 3, and 4. In conclusion a checklist, called a testing protocol, is presented which shows various tasks and questions drawn together by level.

Introduction

This paper describes three types of structuring present in an ideal oral interview, focusing on those tasks and question types which help to insure content validity. Content validity is here defined as the degree to which the oral interview procedure makes possible the elicitation of a speech sample evaluatable in terms of the Foreign Service Institute (FSI) criteria (U.S. Department of State, 1979; Lowe, 1976a). Definitions of the FSI oral proficiency levels are given in the appendix to this paper. Note that S-0 refers to no speaking proficiency and a rating of S-0+ is possible, so that 11 oral proficiency levels from S-0 to S-5 are distinguished.

Characterization. The oral interview has been characterized as a "relaxed, natural conversation," but this strikes me as being wide of the mark in two

respects: everyone knows that the interview is a test (a point Lado [1975:7] stresses); and it is conducted under rather severe time constraints. Whereas a natural conversation might last for several hours, the oral interview is most often completed in ten to thirty minutes. Consequently, I prefer the characterization "conversational interview," which in my mind captures the essence of control over the interview by the interviewer.

Control. How is the interview controlled? At the very least, the interview has a prearranged, deliberate structure. In point of fact, at the Language School (LS, formerly L[anguage] L[earning] C[enter]), three kinds of structure are distinguished: overall structure, specific task structure, and structured question types to elicit information as well as to call forth certain types of task performance.

Overall Structure

At the LS, the oral interview is divided into four phases: Warm-Up, Level Check, Probes, and Wind-Up. The candidate is put at ease with the Warm-Up, has his level of speaking proficiency determined by the Level Check, is pushed beyond this level by the Probes, and is given a feeling of accomplishment with the Wind-Up. A more detailed description of the four phases of the oral interview may be found in Lowe (1976a: 4).

Task structure. Warm-Up and Wind-Up contribute only indirectly to the overall evaluation, so specific task structure normally appears in the two middle phases, the Level Check and the Probes. Notwithstanding the fact that the "conversational interview" is structured in terms of the goal of the interviewer (i.e., determining the candidate's level of speaking proficiency), reaching that goal permits — indeed, requires — considerable flexibility in the specific course set by the interviewer. Were each candidate tested with a rigid format, compromise of the test would be assured, particularly if the specific topics discussed remained the same for all candidates. How, then, is this barrier overcome?

Question-type structure. When I started to work at the LS four years ago I surveyed the staff, asking them what questions were most effective in eliciting a ratable sample of oral behavior. The survey led to the discovery that specific topics of conversation could be changed, but that the general question types could remain the same from test to test, with little if any effect on content validity and no test compromise. For example, "What would you do if you had all the money you ever needed?" could be asked in one test, while "What would you like to accomplish if you were an astronaut?" could be asked in a second test. The specific topics are irrelevant to the goal of the interview; the type of question — hypothetical questions — is the key. Elaboration of this theme may be found in Lowe (1976b).

During the course of the LS survey we also discovered that certain question types were most useful at specific levels of proficiency, while others had a much

wider range of application. Furthermore, many question types were shown to lend themselves especially well to elicitation of specific task behaviors. For example, Polite Requests have been shown to elicit Descriptions and Narrations at Level 2. By assuring that appropriate question types and tasks are included in each test for specific levels, content validity can be much improved.

I will now describe five of the 11 FSI oral proficiency levels (0+, 1, 2, 3, 4), and then discuss how the components of the LS elicitation technique (tasks, functions, and question types) may be drawn together into a checklist to remind the interviewer what characteristics of speaking behavior at each proficiency level need to be addressed. The end result will be a criterion-referenced test with the performance criteria specified at each proficiency level.

Level-by-level description

Whatever constraints may be placed upon the testing scenario, conversation is still basic to the oral interview. The complexity of the conversation, and hence the ultimate nature of the interview, will, of course, reflect the candidate's level of proficiency. This will be seen in the following descriptions.

Level 0+. Conversation at the Level 0+ is at a minimum: it may be virtually non-existent. Typically, the candidate can produce the first few lines of a beginning "dialog." After the initial exchange, however, he is apt to grope for words and to abuse grammar, and the interviewers (our oral interviews are conducted with two interviewers per candidate) may have difficulty coaxing more out of him. In any event, once this point is reached (with care being taken not in any way to embarrass the candidate), checks can be made on several of the 0+ subject areas to determine if the candidate has been exposed to the language and if he commands at least a modicum of control over it. The 0+ Level subject areas are as follows:

basic objects	family members	weather
basic colors	months	weekdays
clothing	time	year
day's date		

The candidate can achieve 0+ in any one of three ways: after the initial exchange ("How are you?", etc.), he demonstrates an ability to carry on a truncated conversation using his limited vocabulary; or he fully answers two or three of the 0+ subject areas, such as naming all of the months or all of the days of the week; or he provides fragmented answers to four or more such ubject areas (two months and three days of the week plus the time and the weather, for example). At the LS, the information acquired from an oral interview at this level is most likely to be used in placing a candidate in an ongoing introductory class where his previous knowledge will allow him to catch up to the other students. I envision a similar use for the oral interview in an academic setting.

Level 1. Although the candidate can create original sentences and phrases,

which a 0+ usually cannot, conversation at this level may leave a great deal to be desired. The candidate can function in a question-and-answer mode, usually reserving the role of respondent for himself. If this occurs, the interviewers can ask the candidate to pose some questions, thus checking for the two ingredients necessary for any full conversation. Because Level 1 is regarded as a survival level of proficiency, ascertaining if a candidate can ask questions is probably more crucial at this level than at any other. We believe that this ability can be assumed at Level 1+ or higher, although we are not certain to what extent one linguistic behavior can be inferred from the presence of another; this point ought to be investigated separately.

Similarly, Basic Situations (Lowe, 1976b: 13) must also be checked to determine if the candidate can survive on his own (the basic requirement of Level 1 proficiency). The question is not how accurately he performs, but how effectively he communicates through the use of his target language behavior.

Level 2. Here we look beyond sheer survival behavior for the added ability to describe and narrate (narration being a more complicated task which includes description). Recall that Polite Requests can elicit material suitable for evaluating such performance. Along with these general abilities, we further expect use of non-present times (past and future in some form). If the candidate is a weak 2, we may ask him to carry out some Level 1 tasks, such as Basic Situations, in order to assure ourselves that he is not a Level 1 or 1+ speaker. If he is a strong 2, we may ask him to attempt some Level 3 tasks.

Level 3. This is the level of Minimum Professional Competence — crucial in government work because it is the target level for many overseas assignments. It differs qualitatively from the levels below it because Level 3 speakers evince a fluency which clearly surpasses that at the lower levels. The Level 3 speaker controls general vocabulary to the extent that he need not grope for words (although uncertainty with a particular technical vocabulary is still to be expected). His basic grammar is handled with assurance and with few errors; more complex grammar will often cause problems. He is expected to treat unkown topics and situations while not losing control of his grammar or his vocabulary.

Three tasks are particularly effective at Level 3: Unknown Topics, Unknown Situations, and Supported Opinion. By "unknown" I mean that the candidate probably has not had a previous opportunity to address the topic or situation in the target language, although he may well have dealt with them in his native language.

For an Unknown Situation we usually give the candidate written instructions in English and ask him to roleplay in the target language with one of the interviewers. For example:

> You are in a western European country on a superhighway when you have a *blowout*. Luckily, you are near an emergency phone.
>
> Call for help, explaining that you have an *oddsized, tubeless* American tire, you need to replace it, and you also need a *tow truck* to help you out of the *ditch* you landed in when the tire blew.

> We realize that you may not have the exact vocabulary for this situation, but do the best you can to make yourself understood.

Some candidates do not like to roleplay, but for most, Unknown Situations is the technique of choice.

Unknown Situations have the following advantages: they are short and precise; they allow the interviewers to expose vocabulary, grammar, and sociolinguistic problems; and they permit the testing of reality in an artificial environment where the abstract might otherwise be dominant. But a word of caution — Unknown Situations should not become "interpreter situations" (see Clark, 1972: 121), which are often too long, and which can rob the interview of badly needed time.

Supported Opinion is another technique which can be used effectively at Level 3 or higher. A Descriptive Prelude or Conversational Prelude can introduce the topic and the candidate can then be required to elaborate on the theme. For example:

> "You are undoubtedly aware of the struggle between the automotive industry and advocates of public transportation." ("Yes.") "Which form of transportation would you promote and why?"

This line of questioning allows the interviewers to set the stage linguistically and to shift levels stylistically if need be. Of course, there is no guarantee that the candidate will shift levels along with them, but it is worth the effort. As for content validity, the interviewers have given the candidate a chance to express Supported Opinions. In the event that the first topic fails to work, the interviewers may elect to try a second or even a third — provided that the candidate isn't traumatized by the situation, and that there is sufficient time left in the interview.

Level 4. Jones (1978: 89) is correct: this level is difficult to deal with because the FSI proficiency definitions do not adequately distinguish Level 4 from Level 3. In any event, the candidate is expected to employ a precise and extensive vocabulary and to tailor his speech to the sociolinguistic environment. Once again, the Unknown Situation is a useful interview technique to determine the candidate's speaking proficiency level.

One of the most important characteristics of the Level 4 speaker is his ability to scale his language to the level of the person to whom he is speaking. This ability can be tested by using an Unknown Situation in which the candidate roleplays with someone who is not on his sociolinguistic level. For example, the candidate assumes the role of a tenant in a fashionable, high-rise apartment and is instructed to report a plumbing problem (for example) to the building superintendent. In American English, it would be "poor form" were the tenant to open the conversation by saying "Oh, Mr. Building Superintendent, . . ." Adult Americans in such a situation would probably address him as "Super" or by his first name. This is only one of a number of illustrations which could be cited. Whatever the technique, the ability to tailor the language to the particular situa-

tion determines, in large measure, whether or not the candidate has achieved Level 4 proficiency. It should be noted that Jones (1978) has published a series of higher-level probes.

Synopsis. The preceding discussion suggests that it should be possible to administer oral interviews characterized by rather specific task and question type structure across a series of interviews for the same proficiency level. The intent of doing so, of course, is to increase content validity. One mechanism to insure such uniformity is to use a testing protocol, such as that discussed in the next section.

Testing protocol

The testing protocol (Figure 1) progresses by levels (0+ through 5) via a series of items usually unique to each proficiency level. Properly filled out, the protocol should have most (or all) of the boxes checked at the candidate's proficiency level, with some boxes checked at the next higher level in order to make sure that Probes have been attempted. (At Levels 1 and 2, some particularly useful probes are listed along with the obligatory checks.) Thus, a Level 3 protocol would have most of the Level 3 boxes checked, along with several at the next higher Level (such as "Placed him in unfamiliar situations and topics" and "checked for supported opinion," as cases in point).

Like the Proficiency Definitions from which it is derived, the protocol combines both structured tasks and functions. (For a discussion of this problem, see Clark, 1972: 126.) Moreover, the protocol contains question types which have proven to be useful at specific proficiency levels. Thus, it is possible to draw the major strands together in one checklist, and by use of such a list to strengthen content validity.

FIGURE 1
Testing Protocol

LEVEL 0+: Tried to have conversation? _____

Covered 0+ Subject Areas: Which?

Basic objects	_____	Months	_____
Basic colors	_____	Time	_____
Clothing	_____	Weather	_____
Day's date	_____	Weekdays	_____
Family members	_____	Year	_____

LEVEL 1: Tried to have conversation? _____

Checked for minimum courtesy requirements? _____

Checked that he can handle simple situations of daily life and travel (S-1 Situations)? _____

Had him ask you questions? _____

Tried props when conversation fails? _____

Probed for past tense(s) and future? _____

LEVEL 2: Checked how he can satisfy routine social demands? _____

Checked how he talks about autobiographical information? _____

Checked how he talks about current events? _____

Checked how he uses basic structures? _____

Checked how he uses more complex structures? _____

Checked for description? _____

Checked for narration, particularly in past & future? _____

Checked how he handles simple situations of daily life and travel (S-1 Situations)? _____

Checked how he joins sentences in connected discourse? _____

Probed for how he handles an unknown topic or situation? _____

Probed for supported opinion? _____

LEVEL 3: Checked both everyday and abstract subject matter? _____

Placed him in unfamiliar situations and topics? _____

Checked his control of grammar? _____

Checked for supported opinion? _____

Checked for description? _____

Checked for narration? _____

Checked how he uses low-frequency structures? _____

Checked how he uses complex structures? _____

Checked for broad vocabulary? _____

Checked how he answers hypothetical questions? _____

LEVEL 4: Checked both everyday and abstract subject matter? ____

Placed him in unfamiliar situations and topics? ____

Checked his control of grammar? ____

Checked for supported opinion? ____

Checked for description? ____

Checked for narration? ____

Checked how he uses low-frequency structures? ____

Checked how he uses complex structures? ____

Checked for broad vocabulary? ____

Checked for how he answers hypothetical questions? ____

Checked how he handles an unknown situation? ____

Checked how he tailors his speech to his audience(s)? ____

LEVEL 5: Checked both everyday and abstract subject areas? ____

Checked for high-level colloquialisms? ____

Checked for pertinent cultural references? ____

Checked his ability to converse freely and idiomatically in his special fields? ____

Checked that he speaks and sounds like an educated native speaker in all that he says? ____

Checked how he handles unknown situations and topics? ____

REFERENCES

Clark, John L. D., 1972. *Foreign language testing: theory and practice*. Philadelphia: The Center for Curriculum Development.

Foreign Service Institute, 1979. *Testing kit: French and Spanish*. Washington, D.C.: U. S. Department of State.

Jones, Randall L. 1978. Interview techniques and scoring criteria at the higher proficiency levels. In John L. D. Clark, ed. 1978. *Direct testing of speaking proficiency: theory and application*. Princeton, N.J.: Educational Testing Service.

Lado, Robert, 1975. Comments in Randall L. Jones and Bernard Spolsky, eds., *Testing language proficiency*. Arlington, Va.: Center for Applied Linguistics.

Lowe, Pardee, Jr. 1976a. *The oral language proficiency test*. Washington, D.C.: U.S. Government Interagency Language Roundtable.

_____. 1976b. *Handbook on question types and their use in LLC oral proficiency tests* (preliminary version). Washington, D.C.: Central Intelligence Agency.

APPENDIX
Absolute Oral Language Proficiency Ratings
(from Foreign Service Institute, 1979: 13-15)

As currently used, all the ratings except the S-5 may be modified by a plus (+), indicating that proficiency substantially exceeds the minimum requirements for the level involved but falls short of those for the next higher level.

Elementary proficiency

S-1 *Able to satisfy routine travel needs and minimum courtesy requirements.* Can ask and answer questions on very familiar topics; within the scope of very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak the language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at the S-1 level should be able to order a simple meal, ask for shelter or lodging, ask for and give simple directions, make purchases, and tell time.

Limited working proficiency

S-2 *Able to satisfy routine social demands and limited work requirements.* Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on nontechnical subjects (i.e. topics which require no specialized knowledge) and has a speaking vocabulary sufficient to respond simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.

Professional proficiency

S-3 *Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics.* Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary

is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.

Distinguished proficiency

S-4 *Able to use the language fluently and accurately on all levels normally pertinent to professional needs.* Can understand and participate in any conversation within the range of own personal and professional experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

Native or bilingual proficiency

S-5 *Speaking proficiency equivalent to that of an educated native speaker.* Has complete fluency in the language such that speech on all levels is fully accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

Section II

Empirical Research

A Study of the Reliability and Validity of the Ilyin Oral Interview

Alice Engelskirchen,
Elinore Cottrell, and John W. Oller, Jr.
University of New Mexico

Abstract. Reliability and validity of the Ilyin Oral Interview (IOI) are examined with respect to interscorer agreement. Interviews of 11 students from an ESL class at the University of New Mexico were taped and later scored by 20 native speakers of English. All scorers were either practicing ESL teachers or ESL teachers in training. Interscorer agreement in the IOI scores showed a 79% variance overlap across the 12 most consistent judges and a 45% variance overlap across scores and external validity criteria. The latter included ratings of the IOI interviews by two judges on the five FSI Oral Interview scales and an independent ranking of the 11 students interviewed by their regular ESL teacher. Item analysis included a questionnaire assessing the pragmatic appropriateness of the questions in the IOI. Interscorer agreement shows the IOI to be a dependable measure of oral proficiency even in the case of relatively homogeneous ability levels and with minimal instructions to scorers. Items which the scorers felt were more natural were generally better discriminators.

Until recently, oral proficiency was probably the least studied area of language testing. However, in the last few years major attention has been focused on this topic. At least one entire conference was devoted to oral testing in 1978 (Clark, 1978) and there is now a three-year-old newsletter devoted to interview testing (Lowe, 1976-79). Perhaps the lack of research can be attributed to the difficulty of administering and scoring oral tests. Any such research is costly; one-to-one interviews are extremely time consuming; and scoring is difficult because of the fleeting nature of the spoken word. Taping introduces additional time requirements and a possible need for transcription, and technical quality of recordings immediately becomes an issue.

In the face of such difficulties, the need for reliable and valid methods of assessing oral proficiency seems clear. The Ilyin Oral Interview (Ilyin, 1972,

1976) is one attempt to fill the gap. Although Mattran (1978) sees the IOI as a discrete-point test, it is actually a kind of compromise between discrete-point and integrative testing. It attempts to relate certain structures of English to common contexts of communication in a pragmatically viable way. While the IOI does not distinguish the time-honored components of oral proficiency (pronunciation, vocabulary, grammar, fluency, and comprehension) as is done in the Foreign Service Institute scales, its solution of attempting to measure global oral proficiency is well supported in the current research literature. Work by Scholz et al. (1979), Hendricks et al. (1979), Callaway (1979), and Mullen (1977) shows that a single global factor accounts for the bulk of reliable variance in all of the scales so far investigated. In fact, it can be argued that both trained and naive judges seem equally incapable of distinguishing the various characteristics of speech that the multiplicity of scales aim at. They seem good at judging one central variable — probably it should be called "communicative effectiveness."

The IOI is a structured questionnaire based on a sequence of pictures depicting common events in the daily life of a student. Figure 1 displays the pictures that are used during the orientation part of each interview.

From left to right each set of pictures represents a sequence of activities in the life of a certain fictitious character either on a weekend or a weekday. Days are indicated in the upper left hand corner of each sequence and times of each activity are indicated in the clock (or clocks) under each picture. Questions put to the examinee pertain to the activities pictured. For instance, after an introduction to the principal character (in this case, Bill) and to the general format of the test, the examinee might be asked questions such as "What does Bill do every evening from 9:30 to 10:00?"

Scoring of responses is based on a three-point scale (0, 1, or 2), indicating "no response", "unintelligible response", or "inappropriate response"; "appropriate and intelligible response with one or more grammatical errors"; and "appropriate and grammatical response", respectively.

While a number of empirical studies have indicated substantial reliability and validity of the IOI technique, no specific study of interscorer agreement has been reported in the published literature. We will argue that the sort of agreement that is required across native judges is not only a prerequisite reliability criterion, but is actually the most appropriate validity criterion for any such test. We asked: (1) To what extent do native judges (with a minimum amount of training) arrive at similar scores on the IOI? (2) How much agreement is there across IOI scores and global ratings of examinees? (3) Will the correspondence between question and picture or the pragmatic naturalness of the questions affect the estimated validities of items on the IOI?

Method

Eleven foreign students enrolled in English 103 (fairly advanced ESL learn-

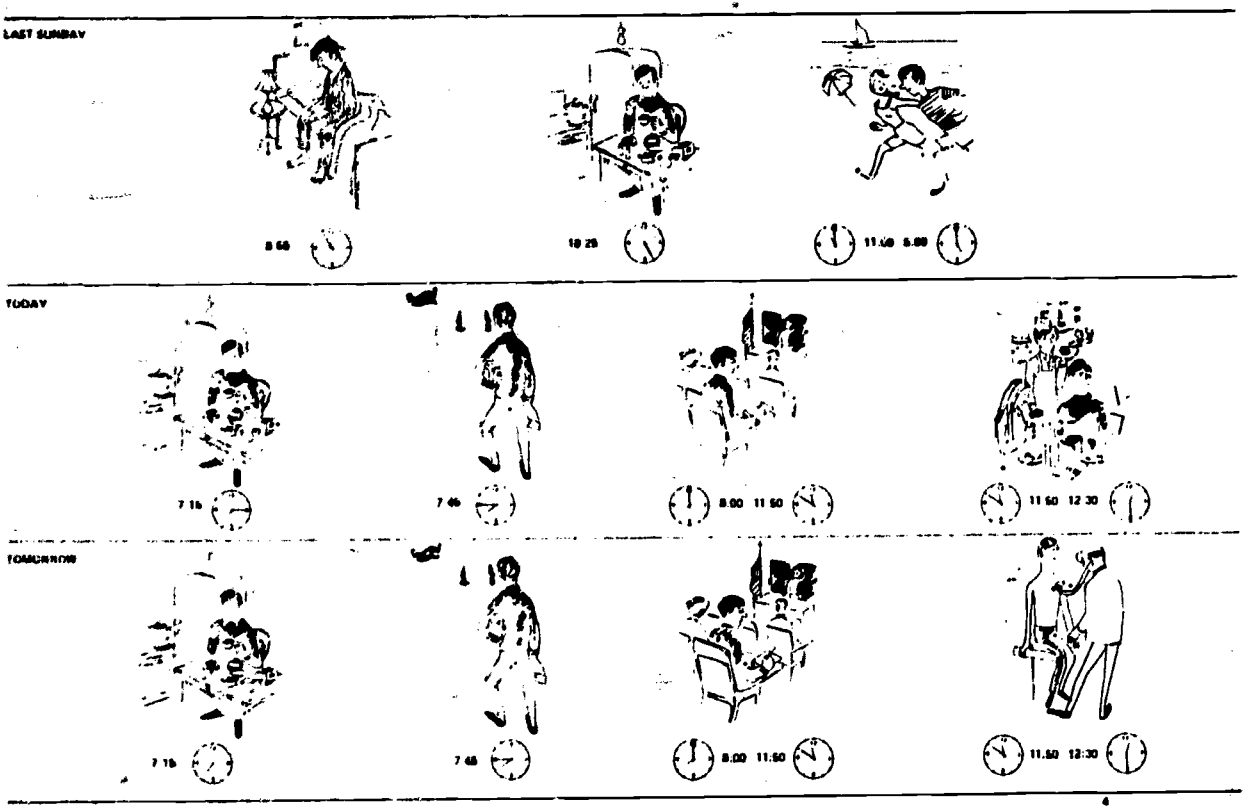


Figure 1. Orientation pictures from the Ilyin Oral Interview (1976)

ers) at the University of New Mexico were interviewed using the IOI (30-item, short form, Bill). Students' native languages included Spanish, Japanese, Indonesian, Persian, Mandarin, and Finnish. Each interview was recorded on a portable cassette tape recorder (Superscope, C-103A). The same machine was used for playback.

Twenty language teachers or teachers in training listened to each of the recorded interviews and scored the responses. The tapes were scored by people working in groups on two occasions,[1] or by people working in pairs at home.[2]

External validity criteria against which the IOI scores were correlated included (1) a global rating of the intelligibility of each subject on a five-point scale by each of the 20 scorers; (2) a ranking of the 11 examineees by Thomas E. Beck, their regular classroom teacher; and (3) ratings on five FSI-type scales of pronunciation, comprehension, fluency, grammar, and vocabulary of each of the 11 examinees separately done by the first two authors of this paper.

Results and Discussion

To determine the degree of agreement across native judges concerning the IOI scores, the 20-by-20 correlation matrix was factored to a single principal component solution. The results are given in Table 1. Loadings (or correlations) with the general factor can be read as indices of the amount of agreement that exists across judges. Since all the estimates taken together indicate the consensus of the 20 judges, the tendency to agree with that consensus can be read as a validity coefficient for each of the judges taken singly, or the overall agreement (the average loading) can be taken as an index of the overall validity of the IOI. Putting it differently, if native speaker consensus is a reasonable validity criterion, loadings on the general factor displayed in Table 1 can be read directly as indicators of test validity.

It can be seen immediately that some judges tend to agree with the consensus more than others. Judges 12 and 13, for instance, showed the lowest degrees of correspondence, while judges 1, 3, 16, and 20 showed quite high agreement with the general consensus. If we consider the 12 most consistent judges only, the average loading is .89, which indicates a variance overlap of 79% across these judges. If we take all 20 into account, the average loading is .81, with a variance overlap across all judges of 66% of the total variance in all the scores. Either way we look at it, the test shows substantial validity and we may infer that it has even higher reliability. It is worth noting at this point too that the

[1]Group 1 included Tomas Buchart, Elinore Cottrell, Charles Decker, Kathy Faulstick, Michael Hays, Karen Jackson, Suzanne Leibundguth, Arthur Maes, Ruth Mindell, John Oller, David Sperow, and Les Wilkin. Group 2 included Sandra Hoogerwerf, Teri McKeigan, Kris Olson, and Neddy Vigil. Nonnative speakers participating in the rating were Farida Kahn, Ingrid Klepper, Hooshang Mehrnoosh, Hans-Dieter Mittendorf, Luis Veléz, and Ryuichi Yorozuya, although data from their participation was not included in the study.

[2]Ratings were done at home by Barbara and Dennis Muchisky and by Ingrid and Stanley Burg.

cards were stacked against the IOI by the poor quality of the tape recordings, by the minimal training of judges, and by the relative homogeneity of the 11 subjects interviewed.

TABLE 1
Principal Component Solution
Revealing Loadings on a Global Proficiency Factor
for Scores of 11 Foreign Students on the Ilyin Oral Interview
as Scored by Native Speakers of English

Judge	Loading on g
Scores by Judge 1	.95
Scores by Judge 2	.83
Scores by Judge 3	.93
Scores by Judge 4	.80
Scores by Judge 5	.78
Scores by Judge 6	.87
Scores by Judge 7	.90
Scores by Judge 8	.79
Scores by Judge 9	.85
Scores by Judge 10	.84
Scores by Judge 11	.85
Scores by Judge 12	.66
Scores by Judge 13	.62
Scores by Judge 14	.60
Scores by Judge 15	.85
Scores by Judge 16	.95
Scores by Judge 17	.89
Scores by Judge 18	.68
Scores by Judge 19	.68
Scores by Judge 20	.96

Mean loading = .81
Total variance accounted for = .66

The second question concerned the correspondence of IOI scores with independent global ratings of the same examinees. In Table 2 we display a principal factor solution for the scores assigned by the 20 judges along with the global intelligibility ratings assigned by the same 20 judges. Actually, the ratings here are not truly independent of the scores since they were assigned immediately after the scoring had taken place. Presumably the scores were still fresh in the minds of the judges and might be expected to influence the ratings. However, as can be seen by a careful examination of Table 2, the scores on the whole proved to be more valid measures of the consensus of scores and ratings than were the ratings. The average loading of scores on the general factor was .76 while the average loading of ratings was .51. The proportion of variance in the general

factor explained by IOI scores is more than twice as large as the proportion accounted for by ratings (58% versus 26%). We may conclude that the IOI scoring system is superior to a simple rating of degree of intelligibility. Scores and ratings taken together produce an average loading of .64 or a total common variance of 41%.

TABLE 2
Principal Component Solution
Revealing Loadings on a Global Proficiency Factor
for Scores of 11 Foreign Students on the Ilyin Oral Interview
and Ratings of their Intelligibility by Native Speakers of English

Judge	Loading on g	Judge	Loading on g
Score by Judge 1	.88	Rating by Judge 1	.68
Score by Judge 2	.88	Rating by Judge 2	.71
Score by Judge 3	.93	Rating by Judge 3	.73
Score by Judge 4	.83	Rating by Judge 4	.66
Score by Judge 5	.71	Rating by Judge 5	.66
Score by Judge 6	.85	Rating by Judge 6	.87
Score by Judge 7	.87	Rating by Judge 7	.53
Score by Judge 8	.70	Rating by Judge 8	.78
Score by Judge 9	.85	Rating by Judge 9	.84
Score by Judge 10	.83	Rating by Judge 10	.63
Score by Judge 11	.89	Rating by Judge 11	.64
Score by Judge 12	.47	Rating by Judge 12	.27
Score by Judge 13	.45	Rating by Judge 13	.34
Score by Judge 14	.51	Rating by Judge 14	.17
Score by Judge 15	.83	Rating by Judge 15	.75
Score by Judge 16	.92	Rating by Judge 16	.02
Score by Judge 17	.85	Rating by Judge 17	.66
Score by Judge 18	.65	Rating by Judge 18	.12
Score by Judge 19	.73	Rating by Judge 19	.35
Score by Judge 20	.88	Rating by Judge 20	.28
Mean loading for scores = .76		Mean loading for ratings = .51	
Variance accounted for in scores = .58		Variance accounted for in ratings = .26	
Mean loading overall = .64			
Total variance accounted for in scores and ratings = .41			

In Table 3 the general factor solution for scores and the external criteria are given. Here, the ratings by Beck were based on extensive classroom interaction with the 11 examinees in question. The FSI type ratings, by contrast, were largely based on the IOI interviews themselves.

Again the loadings for scores were generally higher than those for ratings. The average for the former was .79 while for the latter it was .67. The contrast is more marked, of course, if we consider the amount of variance explained by

TABLE 3
Principal Component Solution
Revealing Loadings on a Global Proficiency Factor
from Scores on the Ilyin Oral Interview and Independent Ratings

Judges	Loading on g
Scores by Judge 1	.89
Scores by Judge 2	.87
Scores by Judge 3	.92
Scores by Judge 4	.76
Scores by Judge 5	.66
Scores by Judge 6	.82
Scores by Judge 7	.88
Scores by Judge 8	.79
Scores by Judge 9	.86
Scores by Judge 10	.88
Scores by Judge 11	.82
Scores by Judge 12	.55
Scores by Judge 13	.55
Scores by Judge 14	.45
Scores by Judge 15	.79
Scores by Judge 16	.92
Scores by Judge 17	.94
Scores by Judge 18	.69
Scores by Judge 19	.76
Scores by Judge 20	.93
E.C. Comprehension	.32
E.C. Fluency	.50
E.C. Grammar	.82
E.C. Vocabulary	.79
E.C. Pronunciation	.78
A.E. Comprehension	.52
A.E. Fluency	.69
A.E. Grammar	.65
A.E. Vocabulary	.81
A.E. Pronunciation	.78
Beck Rank Order	.74

Mean loading of scores and independent ratings = .75
Variance accounted for = .56

Mean loading of scores = .79
Variance accounted for = .62

Mean loading of independent ratings = .67
Variance accounted for = .45

each of the types of measures. Scores accounted for 62% of the total variance in the global proficiency factor, and ratings accounted for only 45% of the total. Considering the limitations noted earlier and the fact that the subjects interviewed were at a relatively homogeneous level of ability due to the placement procedure by which they were assigned to Mr. Beck's 103 class, both scores and

ratings seem to have substantial reliability and validity. However, as we found in reference to Table 2 above, the scores seem to produce a greater amount of valid variance than do the more subjective ratings.

We now come to the third question posed earlier. What is the effect of the correspondence between question and picture and of the pragmatic naturalness of the questions on estimated validities of items? To answer this question we did a rather special sort of item analysis. In addition to the standard item facility and item discrimination indices, recorded in columns 1 and 2 of Table 4, we asked the same 20 judges who did the scoring of interviews to rate each item on two five-point scales. The first scale concerned the fit between picture and question. The issue was whether or not the question seemed to make sense in relation to the pictured event or situation. The question judged lowest in degree of fit was the one that asked, "If Bill were on a bus, what would he be doing?" This item seemed odd because there is no obvious basis for inferring why Bill might be on a bus in the first place. The very idea of the bus seems unmotivated by the pictures. The second scale concerned the naturalness of the question itself. A question which was judged low in naturalness involved the instruction, "Ask a question about this picture with the word 'if'." Results for the picture fit and naturalness scales are given in columns 3 and 4 of Table 4.

TABLE 4
Item Analysis for the Ilyin Oral Interview (Bill, short form)

Item number	Item facility	Item discrimination	Picture fit	Naturalness	Negative points
4. What time does Bill usually study? (1)	.56	.28*	4.65	4.40	1
5. What does he usually do every evening from 9:30-10:00? (2)	.59	.51	4.80	4.25	0
8. How does he go to school? (3)	.68	.51	4.00	3.75*	1
9. Where does he eat lunch on weekdays? (4)	.62	.16*	3.75*	3.85*	3
10. When does he eat lunch on weekdays? (5)	.69	.30*	4.40	4.05	1
11. Is he going to be eating lunch tomorrow at 12:15? (6)	.66	-.09*	4.30	4.10	1
12. What will he do tomorrow at 12:15? (7)	.59	.42	4.05	3.55*	1
13. What is Bill going to do today before he watches TV? (8)	.59	.50	4.60	4.25	0
15. When will dinner be eaten tomorrow? (9)	.76	.11*	4.00	3.10*	2
18. Where did he go with a girl on Sunday? (10)	.77	.39	4.45	3.80*	1

ratings seem to have substantial reliability and validity. However, as we found in reference to Table 2 above, the scores seem to produce a greater amount of valid variance than do the more subjective ratings.

We now come to the third question posed earlier. What is the effect of the correspondence between question and picture and of the pragmatic naturalness of the questions on estimated validities of items? To answer this question we did a rather special sort of item analysis. In addition to the standard item facility and item discrimination indices, recorded in columns 1 and 2 of Table 4, we asked the same 20 judges who did the scoring of interviews to rate each item on two five-point scales. The first scale concerned the fit between picture and question. The issue was whether or not the question seemed to make sense in relation to the pictured event or situation. The question judged lowest in degree of fit was the one that asked, "If Bill were on a bus, what would he be doing?" This item seemed odd because there is no obvious basis for inferring why Bill might be on a bus in the first place. The very idea of the bus seems unmotivated by the pictures. The second scale concerned the naturalness of the question itself. A question which was judged low in naturalness involved the instruction, "Ask a question about this picture with the word 'if'." Results for the picture fit and naturalness scales are given in columns 3 and 4 of Table 4.

TABLE 4
Item Analysis for the Ilyin Oral Interview (Bill, short form)

Item number	Item facility	Item discrimination	Picture fit	Naturalness	Negative points
4. What time does Bill usually study? (1)	.56	.28*	4.65	4.40	1
5. What does he usually do every evening from 9:30-10:00? (2)	.59	.51	4.80	4.25	0
8. How does he go to school? (3)	.68	.51	4.00	3.75*	1
9. Where does he eat lunch on weekdays? (4)	.62	.16*	3.75*	3.85*	3
10. When does he eat lunch on weekdays? (5)	.69	.30*	4.40	4.05	1
11. Is he going to be eating lunch tomorrow at 12:15? (6)	.66	-.09*	4.30	4.10	1
12. What will he do tomorrow at 12:15? (7)	.59	.42	4.05	3.55*	1
13. What is Bill going to do today before he watches TV? (8)	.59	.50	4.60	4.25	0
15. When will dinner be eaten tomorrow? (9)	.76	.11*	4.00	3.10*	2
18. Where did he go with a girl on Sunday? (10)	.77	.39	4.45	3.80*	1

Item number	Item facility	Item discrimination	Picture fit	Naturalness	Negative points
21. How long did he play cards? (11)	.72	.19*	4.50	4.55	1
22. These questions are in the past. Ask a question about this picture. (12)	.74	.26*	4.35	3.70*	2
23. These questions are about weekdays. Ask one question about these two pictures (13)	.53	.02*	3.75*	3.65*	3
25. You have seen many pictures of Bill on weekdays. What does Bill do? (14)	.61	.25*	3.90*	3.90*	3
27. Where had Bill been before he went to the beach last Sunday? (15)	.50	.11*	3.60*	3.70*	3
28. Tell what kind of breakfast he has every morning. (16)	.79	.35	2.45*	3.15*	2
29. Tell what he wears at school. (17)	.72	.37	2.65*	3.35*	2
30. Tell how long he was at the beach on Sunday. (18)	.74	.26*	4.45	3.45*	2
31. Now ask where he went after that. (19)	.66	.38	4.00	3.65*	1
32. Ask who(m) he eats lunch with. (20)	.72	.32	4.15	3.55*	1
36. Ask a question about the big picture with "after."(21)	.61	.16*	4.10	3.35*	2
37. Answer your question. (22)	.73	.13*	4.00	3.37*	2
40. Ask a question about this picture with the word "if."(23)	.52	.39	3.60*	3.10*	2
41. Answer your question. (24)	.60	.30	3.63*	3.11*	2
44. If it were tomorrow at this time, what would Bill be doing? (25)	.62	.04*	3.90*	3.50*	3
45. If Bill were on a bus, what would he be doing? (26)	.41	.19*	1.53*	2.20*	3
46. If it were Sunday at this time, what would Bill have been doing? (27)	.49	.20*	3.30*	3.40*	3
47. What would he have done before that? (28)	.55	.32	3.30*	3.10*	2
48. If he had been sick, would he have gone to the beach on Sunday? (29)	.79	.13*	2.94*	3.85*	3
49. What might he have done? (30)	.60	.23*	3.89*	3.45*	3

*Minimum standards of acceptability were somewhat arbitrarily established as follows: Item facility .85-.15, Item discrimination .30, Picture fit 4.00, and Naturalness 4.00. Asterisks indicate indices below the respective level of acceptability.

In the 5th column of the table, we give the number of negative points assigned to each item on the basis of the four criteria considered: item facility, item discrimination, picture fit, and naturalness. (Item numbers at the extreme left correspond to the full 50-item version of the Bill form.)

The ratings for item facility of all 30 items fell between .15 and .85, which is considered an appropriate range. In other words, they would be judged to be suitable on the whole in difficulty level for the 11 examinees tested. Asterisked items in the next column (under item discrimination) are those which fell below the acceptable level arbitrarily set at a standard of .30. For picture fit and naturalness, we arbitrarily considered items falling below a mean of 4.00 on either scale to be somewhat questionable. Eighteen of the 30 items failed to meet the .30 standard for item discrimination. Further, 14 items fell below 4.00 on both picture fit and naturalness. Of the latter, fully 71% were also weak discriminators.

We may conclude that the items judged higher in naturalness and picture fit were better discriminators. It follows from this conclusion that the overall discrimination of the IOI might be improved significantly by making the items conform more closely to the pragmatic requirements of communication. However, the IOI as it now stands seems to be a quite reliable and valid measure of language proficiency. This conclusion is doubly remarkable in view of the factors which biased the present study against the IOI and also in light of the shortcomings of certain questions in the IOI. We are encouraged to believe that oral tests of this sort can be refined to extremely high levels of reliability and validity.

REFERENCES

Callaway, Don. 1979. Accent and the evaluation of ESL proficiency. In Oller and Perkins, 1979.

Clark, John L. D., ed. 1978. *Direct testing of speaking proficiency: theory and application*. Princeton, N.J.: Educational Testing Service.

Hendricks, Debby, George Scholz, Randon Spurling, Marianne Johnson, and Lela Vandenburg. 1979. Oral proficiency testing in an intensive English language program. In Oller and Perkins, 1979.

Ilyin, Donna. 1972, 1976. *Ilyin oral interview*. Rowley, Mass.: Newbury House.

Lowe, Pardee. 1976-1979. *Interview Testing Newsletter*. Rosslyn Station, Va.: U.S. Government Interagency Language Roundtable.

Mattran, Kenneth J. 1977. Native speaker reactions to speakers of ESL: implications for adult basic education oral English proficiency testing. *TESOL Quarterly 11*, 4.

Mullen, Karen A. 1977. Rater reliability and oral proficiency evaluations. In James E. Redden, ed., *Proceedings of the First International Conference on Frontiers in Language Proficiency and Dominance Testing, held at Southern Illinois University at Carbondale, April 21-23, 1977.* (Occasional Papers on Linguistics, no. 1.) Carbondale, Ill.: Dept. of Linguistics, Southern Illinois University. (Also in Oller and Perkins, 1979.)

Oller, John W., Jr., and Kyle Perkins, eds. 1979. *Research in language testing.* Rowley, Mass.: Newbury House.

Scholz, George, Debby Hendricks, Randon Spurling, Marianne Johnson, and Lela Vandenburg. 1979. Is language ability divisible or unitary?: a factor analysis of 22 English language proficiency tests. In Oller and Perkins, 1979.

Inter-Rater and Intra-Rater Reliability of the Oral Interview and Concurrent Validity with Cloze Procedure

Elana Shohamy
University of Minnesota

Abstract. An oral interview speaking test and cloze tests were administered to students of Hebrew at the University of Minnesota. The taped interviews were rated by three raters on vocabulary, grammar, pronunciation, fluency, and overall speaking proficiency. Inter-rater and intra-rater reliabilities and concurrent validity of the oral interviews with the cloze tests were calculated. The oral interview rating scale and the raters' training procedures are described. The study also assessed students' attitudes towards the two testing procedures.

Introduction

This paper presents partial results of a study investigating the relationship between an oral interview and a cloze test in Hebrew (Shohamy, 1978), focusing on issues related to the oral interview procedure.

An oral interview was developed for testing speaking proficiency in Hebrew, and the following were investigated: inter-rater and intra-rater reliabilties of the oral interview, and concurrent validity of the oral interview with a cloze procedure.

The findings reported here were a prerequisite for the primary purpose of the study, which was to investigate whether the cloze procedure can be used to predict performance on the oral interview in Hebrew — a prerequisite, inasmuch as reliability is a necessary condition for validity.

The paper first briefly discusses the instruments, the oral interview and the cloze procedure, and describes the sample used in the study, the administration of the tests, and their rating and scoring. Analysis of the data, findings, and conclusions follow.

The oral interview

The oral interview used in the study was adapted from that developed by the Foreign Service Institute as a speaking proficiency test and now widely employed by the FSI and other U.S. government agencies, including the CIA and the Peace Corps. In the adapted oral interview, as in the original, oral proficiency is assessed after a structured informal interview that lasts between 15 and 30 minutes. During the interview, speaking skill is exercised in a face-to-face conversational situation and performance is evaluated based on the ability to use and function in the language, not only on the knowledge of distinct linguistic items.

In the FSI interview, descriptive functional statements define levels of general oral proficiency and/or speaking aspects — vocabulary, grammar, pronunciation, fluency, and listening — on a scale ranging from 0 to 5, where 5 is equal to that of a native speaker. The rating scale used in this study is similar but not identical, being based on a rating scale developed by Clifford (1977) for testing German speaking proficiency.

Clifford's rating scale was constructed from six other instruments (the MLA Teacher Qualification Statement, the rating scale from the MLA speaking test, the general FSI proficiency description, the FSI grid of "Factors in Speaking Proficiency", the FSI supplementary proficiency descriptions, and the CIA supplementary rating). Clifford collapsed the matrices of these instruments into one, validated it, and formed a separate rating scale with six levels (0-5) for rating oral proficiency in terms of grammar, vocabulary, pronunciation, and fluency. The main advantage of using Clifford's instrument was that it allowed rating of speaking in *each* of the speaking aspects. (Also, test retest, inter-rater, and intra-rater reliability figures, which were high, were available.)

Three Hebrew language experts from the University of Minnesota participated in the adaptation of Clifford's rating scale to Hebrew speaking proficiency rating. Since the German rating scale provided mainly functional statements of proficiency (describing what a person can do with the language rather than specific linguistic elements), only minimal changes in the grammar and pronunciation scales were necessary. (The adapted scale is given in Appendix A.)

The cloze procedure

The cloze is a testing procedure in which the examinee is required to resupply letters or words that have been systematically deleted from a continuous text. Scores obtained from cloze tests correlate highly with scores of specific skill tests, and with tests attempting to measure overall proficiency, in several languages (Darnell, 1968; Bormouth, 1962; Gregory, 1966; Hinofotis, 1976; Oller & Conrad, 1971; Toiemah, 1978; Leong, 1972; McLeod, 1974). Based on such correlations, some researchers (Aitken, 1977; Stubbs, 1974; Oller, 1973; Oller,

1978) claim that the cloze procedure can be considered a valid test of overall proficiency.

The cloze procedure has also correlated highly with proficiency tests in Hebrew as a second language. Nir and Cohen (1977) report correlations of up to .92 between a cloze test and a composite score obtained from grammar, listening comprehension, and reading comprehension proficiency tests, supporting a conclusion that the cloze in Hebrew follows patterns similar to those in other languages (Nir et al., 1978).

Two cloze tests were used in this study: one, classified as "easy," selected from a beginning level Hebrew textbook; the other, classified as "difficult," selected from an Israeli women's magazine. The selected texts, of 300 words each, were in modern Hebrew and were not related to a specific subject area which only some of the students might have been familiar with. The *sixth* word deletion rule was chosen for both texts (based on a pilot study conducted to determine the deletion rule which best discriminates among the proficiency levels of the participants), so that each test included 50 deletions. Hebrew vowels were used in both texts. Each blank when filled correctly was assigned one point. Hence the score range of each of the cloze tests was 0-50.

The sample

A sample of 106 University of Minnesota students was selected to participate in the study: 65 students enrolled in Hebrew classes during the spring of 1977, 35 students who had enrolled in Hebrew classes some time before, and 6 native Israeli students. A special effort was made to include students representing all levels of language proficiency.

Tests administration

All tests were administered within a period of six weeks during the spring of 1977. Half of the subjects were administered the cloze procedure first and the oral interview second, and the rest in the reverse order.

The oral interviews lasted from 15 to 30 minutes and were all conducted by the researcher.

The interviews followed the four phases suggested by Lowe (1976): warm-up, level check, probes, and wind-up. Typically, subjects of interest (to the interviewee) were identified in the warm-up phase. It was in these topics that the interviewee was pushed up to or beyond his/her level of performance, at which point the interview entered its wind-up phase.

All interviews were audio-taped and ratings were assigned at a later date.

The cloze test was administered along with an instruction sheet which directed the students to read the whole passage first and only then to fill in the blanks with the one word which seemed the most appropriate within the context

of the passage. Students were also instructed that misspellings would not count as long as the word was recognizable.

Rating and scoring

All the 106 taped interviews were rated by three raters (including the researcher) on grammar, vocabulary, pronunciation, and fluency. Between 20 and 32 tapes of the original 106 were randomly selected (four weeks after all tapes were rated) to be re-rated by each rater.

Inter-rater and intra-rater reliabilities were necessary conditions for investigating the concurrent validity of the oral interview with the cloze procedure. Therefore, special emphasis was placed on the background and training of the raters. The raters were all Hebrew language teachers and highly proficient in the language. They were trained by the researcher (who was previously exposed to conducting and rating of the Peace Corps type of oral interview at the Educational Testing Service in 1976).

The training consisted of a basic training session explaining the background of the oral interview and the use of the Hebrew rating scale. A practice session followed, during which sample tapes (not included in the study's sample) were used and rated independently by each rater. These ratings were then compared and discussed in an attempt to arrive at a uniform rating. Such practice sessions were repeated weekly while the study's taped interviews were being rated.

The cloze test was scored twice: once by the *exact* word method, whereby only the word which was originally deleted from the text was considered correct, and once by the *acceptable* scoring method, whereby any word which was considered contextually and grammatically correct was counted as correct. All such words were validated by language experts.

Analysis of the data

The following analysis items are relevant to this presentation: inter-rater and intra-rater reliabilities of the oral interview; and concurrent validity of the oral interview with the cloze procedure.

The oral interview variables analyzed are: vocabulary, grammar, fluency, and pronunciation as assigned by the raters. In addition, three more variables were computed: total rating — the sum of the ratings of the four aspects; noncompensatory rating — equal to the lowest rating received on any of the four aspects; and a global rating — equal to the noncompensatory rating, plus 0.5 if ratings of two or more other aspects exceeded the lowest rating.

The cloze variables analyzed are: easy cloze exact, easy cloze acceptable, difficult cloze exact, and difficult cloze acceptable. In addition two more variables were computed: combined cloze exact — the sum of the scores of the easy

cloze exact and the difficult cloze exact; combined cloze acceptable — the sum of the scores of the easy cloze acceptable and the difficult cloze acceptable.

Inter-rater reliability. Cronbach alpha was computed to express the inter-rater reliability for 102 cases rated by all three raters. The reliability coefficients were computed for the four speaking aspects (grammar, vocabulary, fluency, and pronunciation) and also for total, noncompensatory, and global ratings.

Intra-rater reliability. Correlations were computed to express the intra-rater reliability. The correlations were computed for those interviews which were rated twice by each rater (32 such interviews for rater S, 25 for rater G, and 20 for rater E).

Concurrent validity. Pearson product-moment correlations were computed to express the concurrent validity. Correlations were computed between the average oral interview ratings (obtained from the three raters) and each of the cloze test scores.

Findings

Inter-rater reliability. Reliability coefficients ranged from .938 on pronunciation to .990 on the total rating of the oral interview. Such reliability indicates very close agreement among the three raters as to the oral interview rating (Table 1).

TABLE 1
Summary Table for Inter-Rater and Intra-Rater Coefficients for the Oral Interview

		Intra-Rater Reliability**		
	Inter-Rater Reliability* Coefficients	Rater S r	Rater G r	Rater E r
Area	N = 102	N = 32	N = 25	N = 20
Total	.9908	.949	.996	.983
N-C	.9825	.806	.978	.917
Global	.9862	.914	.986	.979
Grammar	.9791	.966	1.000	.944
Vocabulary	.9800	.933	.980	.969
Pronunciation	.9374	.634	.972	.841
Fluency	.9695	.879	.970	.909

p < .001
* based on three raters
** based on two occasions

Intra-rater reliability. Correlations between the two ratings of the oral interview by each rater were high for grammar, vocabulary, and fluency, and lower for pronunciation (Table 1).

Concurrent validity. Significantly high correlations were found between the average oral interview ratings and each of the cloze scores. These correlations range from .743 between pronunciation on the oral interview and the easy cloze acceptable to .872 between grammar on the oral interview and the combined cloze acceptable. Pronunciation and fluency yielded lower correlations than grammar and vocabulary (Table 2).

The common variance (R^2), which is a measure of how well performance on one test can predict performance on the other, was as high as .6991 between the oral interview total score and the difficult cloze acceptable (Table 3).

TABLE 2
Summary of Correlation Coefficients Between Cloze Scores and Oral Interview Ratings.

	Easy Cloze Exact	Easy Cloze Acceptable	Difficult Cloze Exact	Difficult Cloze Acceptable	Combined Cloze Exact	Combined Cloze Acceptable
	r	r	r	r	r	r
Total	.810	.812	.820	.836	.850	.856
N-C	.792	.799	.826	.840	.850	.850
Global	.803	.803	.825	.843	.849	.854
Vocabulary	.796	.800	.798	.818	.832	.839
Grammar	.810	.816	.839	.857	.862	.872
Pronunciation	.750	.743	.776	.783	.791	.789
Fluency	.771	.768	.763	.782	.798	.803
p < .001	N = 94		N = 95		N = 91	

TABLE 3
R and R^2 Figures for the Cloze Test and the Oral Interview

Cloze / Oral Interview	Easy Exact		Easy Acceptable		Difficult Exact		Difficult Acceptable		Total Exact		Total Acceptable	
	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
Total	.8100	.6575	.8116	.6587	.8196	.6718	.8361	.6991	.8503	.7223	.8557	.7323

p < .001

Conclusions

The oral interview procedure in Hebrew administered and rated as it was for this study has high intra-rater and inter-rater reliabilities.

The findings also suggest a high concurrent validity with a cloze procedure in Hebrew.

The high concurrent validity of the oral interview with the cloze may be related to the instruction and teaching methods used in the Hebrew language classes: At the University of Minnesota an equal emphasis was placed on the acquisition of all language skills rather than on a specific skill.

The relationship between the raters' training and the inter-rater and intra-rater reliabilities must be further investigated to determine necessary and sufficient conditions for acceptable reliabilities: a framework for basic training must be investigated. The repeated training in terms of extent and frequency must also be determined. Since the administration of the oral interview is subjective in nature, what is the impact of this subjectivity on the validity of such a procedure? Is there a need for a more standardized interview model? If the interviewer is found to be a factor in the validity of the oral interview, what selection criteria should be employed to qualify oral interviewers?

While other aspects of the oral interview procedure in Hebrew remain to be further investigated (for example, test retest reliability), the researcher recommends the use of the oral interview for testing speaking proficiency in Hebrew for Israeli institutions' proficiency and placement tests as well as for U.S. universities where Hebrew is taught.

Not directly related to the topic of the colloquium, but nonetheless of more than marginal importance, is the attitude of the examinee toward the oral interview testing procedure. Analysis of responses to Likert scale questionnaires (Appendix B) and to essay questions are displayed in Figure 1 and Table 4, respectively. These indicate a significant difference between attitude toward the two tests: students significantly favored the oral interview over the cloze procedure.

FIGURE 1
Mean Responses for the Seven Statements on the Two Instruments

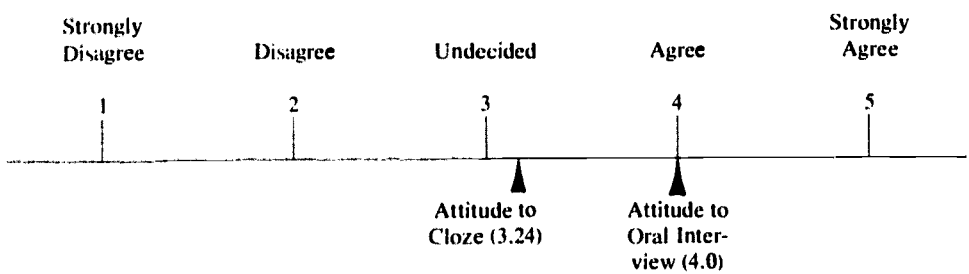


TABLE 4
Frequency and Percentages Based on the
Essay Question on Attitude Toward the Cloze and Attitude
Toward the Oral Interview Procedure

	Attitude Toward the Oral Interview*			Attitude Toward the Cloze**		
	Category	Frequency	Percentages	Category	Frequency	Percentages
Positive	a. Accurate measure of oral ability indicates weak areas.	37	31.89	a. Accurate measure	14	17.95
	b. Like, fun, comfortable	22	18.96	b. Fun, liked, comfortable experience	6	7.69
	c. Helpful, valuable, good opportunity to use language. Need more similar situations	23	19.84	c. Interesting	5	6.41
	d. Interesting, challenging	10	8.62			
	Total positive	92	79.31	Total positive	25	32.05
Negative	a. Made nervous, the tape bothered	14	12.06	a. Difficult, frustrating	28	35.89
	b. Frustrating, difficult	7	6.03	b. Not accurate	12	15.38
	c. Not accurate	2	1.72	c. Couldn't understand, confusing, ambiguous	8	10.26
	d. Disliked	1	.86	d. Disliked	5	6.41
	Total negative	24	20.67	Total negative	53	67.95

*Based on 116 comments. **Based on 78 comments.

REFERENCES

Aitken, K. G. 1977. Using cloze procedure as an overall language proficiency test. *TESOL Quarterly 11*, 1: 59-67.

Bormouth, John R. 1962. Cloze tests as measures of readability. Ph. D. dissertation. Indiana University.

Carroll, John B. 1973. Foreign language testing: will the persistent problems persist? In Maureen Concannon O'Brien, ed. *ATESOL testing in second language teaching: new dimensions*. Dublin, Ireland: The Dublin University Press. 6-17.

Clark, John L. D. 1975. Theoretical and technical considerations in oral proficiency testing. In R. L. Jones and B. Spolsky, eds. *Testing language proficiency*. Arlington, Va.: Center for Applied Linguistics. 10-28.

Clifford, Ray T. 1977. Reliability and validity of oral proficiency ratings and convergent/discriminant validity of language aspects of spoken German using the MLA Cooperative Foreign Language Proficiency Tests (German-speaking) and an oral interview procedure. Ph. D. dissertation. University of Minnesota.

Darnell, D. K. 1968. *The development of an English language proficiency test of foreign students using a clozentropy procedure*. Boulder, Col.: University of Colorado. EDRS: ED 024039.

Davies, Alan. 1978. Language testing. *Language Teaching and Linguistics: Abstracts 11*, 3: 145-59.

Ebel, Robert L. 1967. Estimation of the reliability of ratings. In William A. Mehrens and Robert L. Ebel, eds. *Principles of educational and psychological measurement*. Chicago, Ill.: Rand McNally. 116-31.

Gregory-Panopoulos, J. F. 1966. An experimental application of cloze procedure as a diagnostic test of listening comprehension among foreign students. Ph. D. dissertation. University of Southern California.

Hinofotis, Frances Butler. 1976. An investigation of the concurrent validity of cloze testing as a measure of overall proficiency in English as a second language. Ph. D. dissertation. Southern Illinois University.

Jones, Randall L. 1977. Testing: a vital connection. In June K. Phillips, ed. *The language connection: from the classroom to the world*. The ACTFL Review of Foreign Language Education Series, Vol. 9, Skokie, Ill.: National Textbook Company. 237-65.

Lado, Robert. 1978. Scope and limitations of interview-based language testing: Are we asking too much of the interview? In John L. D. Clark, ed., 1978, *Direct testing of speaking proficiency: theory and application*. Princeton, N.J.: Educational Testing Service. 113-28.

Leong, S. N. 1972. Cloze procedure as a measuring device for reading comprehension in the Chinese language. *NRC 4*. Singapore: Ministry of Education.

Lowe, Pardee, Jr. 1976. *Handbook on question types and their use in LLC oral proficiency tests*. Washington, D.C.: CIA Language Learning Center (preliminary version).

McLeod, J. 1974. *Comparative assessment of reading comprehension: a five-county study: Saskatoon, Canada*. Institute of Child Guidance Development. Mimeograph.

Nir, R. 1974. Hashimush BeShitat HaCloze LeVdikat Shiur HaKriut (The use of the cloze procedure to examine readability). *Iyunim BeHinuch 4*, Sivan: 71-84.

Nir, R. and Andrew Cohen. 1977. Pituach Mivchanei Miyun Lelomdei Ivrit (Development of diagnostic tests for Hebrew learners). Jerusalem: Hebrew University Center for Applied Linguistics. (Paper presented at the 7th International Congress of Jewish Studies, August, 1977.)

Nir, R., Shoshana Blum Kulka, and A. D. Cohen. 1978. *The instruction of the Hebrew language in the Intensive Ulpan in Israel*. Jerusalem: Ruth Bressler Center for Research in Education. Research Report No. 208, Publication No. 578.

Oller, John W., Jr. 1973. Cloze tests of second language proficiency and what they measure. *Language Learning 23*, 1: 105-118.

_____ .1978. Pragmatics and language testing. In Bernard Spolsky, ed., *Approaches to language testing*. Papers in applied linguistics (Advances in language testing series: 2). Arlington, Va.: Center for Applied Linguistics. 39-58.

Oller, John W., Jr., and C. A. Conrad. 1971. The Cloze technique and ESL proficiency. *Language Learning 21*, 2: 183-195.

Oller, John W., Jr., and Frances Butler Hinofotis. 1976. Two mutually exclusive hypotheses about second language ability: factor analytical studies of a variety of language tests. Paper presented at the Annual Meeting of the Linguistic Society of America, Philadelphia, December 1976.

Shohamy, Elana. 1978. An investigation of the concurrent validity of the oral interview with cloze procedure for measuring proficiency in Hebrew as a second language. Ph. D. dissertation. University of Minnesota.

Stubbs, J. B. and R. G. Tucker. 1974. The cloze test as a measure of English proficiency. *Modern Language Journal 58*: 239-241.

Toiemah, Roushdy Ahmed. 1978. The use of cloze to measure the proficiency of students of Arabic as a second language in some universities in the United States. Ph. D. dissertation. University of Minnesota.

APPENDIX A
Hebrew Oral Proficiency Rating Grid

Grammar	Vocabulary	Pronunciation	Fluency
Entirely inaccurate.	Inadequate for even simple conversation	Unintelligible to native speaker	So halting & fragmentary that conversation impossible
Accuracy limited to set expressions: almost no control of syntax; often conveys wrong information. Present tense, simple statements, & question word order.	Limited to familiar topics & to basic personal & survival areas; greetings, time, meals & lodging, purchasing, directions, common expressions.	Frequent gross errors, very heavy accent. Few or no phonemic contrasts. All English sounds. Difficult to understand without repetition.	Speech slow, exceedingly halting, strained, & stumbling except for short or routine sentences and memorized expressions. Difficult to perceive continuity in utterances.

2	Fair control of most basic syntactic patterns: conveys meaning in simple sentences most of time. Some major patterns uncontrolled. Uses correctly, at least sometimes, past & future tenses, conditional, *sh-*, adj. agreement, pronouns, infinitives, & word order.	Adequate for most social situations including introductions, casual conversations about current events, limited work requirements, family, self, daily routine, & hobbies. Expressed simply, with few idioms & with circumlocutions.	Some phonemic inaccuracy, with much allophonic inaccuracy. Foreign accent which requires careful listening. Mispronunciations lead to occasional misunderstanding.	Usually hesitant and jerky. Sentences may be left uncompleted, but he or she is able to keep the conversation going.
3	Limited number of not very serious errors. Imperfect control of some patterns, but always conveys correct meaning. Uses reasonably complex sentences, major word order patterns, correct gender, agreement, and pronoun word order patterns. Correct use of all binyanim.	Sufficient vocabulary to participate effectively in most formal & informal conversations on practical, social, & professional topics; political & social problems, sports, work. Makes frequent & appropriate use of common idioms & colloquialisms.	Identifiable deviations in pronunciation, but with no phonemic errors. Intonation & juncture approximate those of native speaker. Foreign accent evident, occasional mispronunciations occur, but do not interfere with understanding.	Normal rate of speech for most formal & informal conversation, but with some hesitation & unevenness caused by rephrasing and groping for words.
4	Very good command of grammatical structure, & some use of difficult patterns & idioms. Makes only occasional errors, and these show no *pattern* of deficiency.	Professional & general vocabulary broad, precise, & appropriate to the occasion. Can respond appropriately even in unfamiliar situations. Can cope with complex practical & social situations. Large command of idiomatic expressions & colloquialisms.	No consistent or conspicuous mispronunciations, but because of occasional deviations would not be taken for native speaker.	Able to use the language fluently on all levels normally pertinent to professional needs. Participates in any conversation within the range of his experience with a high degree of fluency. Speech effortless & smooth, but non-native in speed & evenness.
5	Performance like an educated native in all ways. Uses difficult & unusual patterns & idioms.	Consistent use of exactly appropriate words. Fully accepted by native speaker.	Native pronunciation. No trace of foreign accent.	Unhesitating and fluent. What pauses there are seem due to search for "right word."

Global Score _____ N-C Rating _____ Global Rating _____

Rater _____ Code Number _____

APPENDIX B
Instrument Assessing Attitude Toward the Cloze* Procedure

Based on the experience of doing the Cloze Tests please indicate your agreement with the following:

1. The testing experience was:	strongly agree	agree	undecided	disagree	strongly disagree
a. comfortable					
b. difficult					
c. unchallenging					
d. fun					
e. pleasant					
f. painful					
g. interesting					
2. I learned a lot from it					
3. It increased my level of confidence in the language					
4. I like this kind of test					

5. Comment in a sentence or two on how you felt about this kind of testing experience.

_____

_____

_____

_____

*An identical instrument was used for the oral interview except that the words "Oral Interview" replaced "Cloze".

*Assessing the Oral Proficiency of Prospective Foreign Teaching Assistants: Instrument Development**

Frances B. Hinofotis,
Kathleen M. Bailey, and
Susan L. Stern

University of California, Los Angeles

Abstract. The language problems of foreign teaching assistants (TA's) at American universities are formidable. At UCLA, the ESL Section of the English Department has responded to the needs of the foreign TA's through the development of an advanced course in oral communication that focuses on teaching-related skills. A research project has been undertaken as well. This paper reports on an outgrowth of the project, the pilot stage in the development of an instrument to be used in assessing the language proficiency of prospective TA's. A panel of raters used the instrument to evaluate video-tapes of students performing a role-play task. Regression analyses were run on the data in an attempt to determine which of the categories on the instrument best predict the overall scores assigned by the raters. In addition to evaluating the subjects on the basis of both global ratings and a series of performance categories, the raters were asked to indicate whether each subject's English was good enough for him to be a TA. On the basis of this study, substantive changes have been effected in the instrument. Further refinements should lead to a performance test of oral proficiency for screening foreign applicants for teaching assistantships.

*This is a revised version of a paper presented at the 1979 Colloquium on the Validation of Oral Proficiency Tests. We wish to thank Roger Bolus, Hossein Farhady, Ebrahim Maddahian, and Mike Bailey for their help with the data analysis, and James Stodel for his technical assistance with the videotapes. We are also grateful to the six raters — Chris Bernbrock, Linda Kimbell, Mike Long, Robert Ochsner, Meredith Pike, and Ann Snow — for their time and interest in the project. In addition, we wish to thank Dr. Andrea Rich, Director of the Office of Instructional Development, for her continuing support of English 34 and the research related to the course.

Introduction

In a 1979 paper entitled "Performance testing of second language proficiency," Randall Jones describes a situation at a large Eastern university where foreign graduate students are employed as teaching assistants (TA's) in disciplines such as chemistry, engineering, mathematics, and psychology. Describing the TA's, Jones says,

> In spite of the fact that they were admitted to the graduate programs and satisfied the English language entrance requirement, some of them cannot be understood by their students, and some have difficulty understanding students' questions and comments (p. 55).

Jones points out that the general ESL tests these students had been given did not measure their English ability in specific situations. Nor did they directly measure speaking proficiency, which Jones calls the skill most critical for teaching.

The situation Jones describes has been noted at a number of universities. In fact, the National Association for Foreign Student Affairs (NAFSA) has recently identified the problems of foreign TA's as a major priority. Communication problems of nonnative speaking TA's have been noted at the University of California at Los Angeles (UCLA) as well. In fact, the oral English proficiency of foreign TA's has been identified as a major problem in undergraduate instruction.

The ESL Section of UCLA's English Department has responded to this problem in two ways. First, it has developed English 34, an advanced course in oral communication for foreign students (Hinofotis and Bailey, 1978; Hinofotis, Bailey, and Stern, 1978). Enrollment priority in this course is given to TA's and graduate students applying for teaching assistantships. Secondly, in connection with this course, research is being conducted to assess the effects of instruction (Hinofotis, Bailey, and Stern, 1979) and provide an instrument to measure the oral English proficiency of foreign students who are applying for teaching assistantships. It is the purpose of this paper to report on the development of that instrument and its use by a panel of raters.

Instrument development

The initial form of the instrument was a checklist which grew out of activities in early quarters of English 34. The purpose of the checklist was to serve as a teaching tool for students to evaluate both their own oral presentations (while viewing themselves on videotape) and those of their classmates. During each oral presentation, the students took notes and then completed the checklist. After class, each student viewed himself on videotape and used the checklist to evaluate his performance. Subsequently, he met with the instructor and they compared their evaluations. Finally, the instructor and the student reviewed the checklists and comments of the other students.

The reactions to the checklist as a teaching tool were mixed. Some students

found it difficult to attend to the content of a speech while trying at the same time to concentrate on so many aspects of the speaker's delivery. Others felt constrained by having to evaluate the speaker in terms of the set categories on the checklist, preferring instead a more open-ended format. However, the major area of discussion was the rating scheme: whether to use numbers (three points, five points, or other), verbal descriptors, or both. The last area became a topic of heated debate in the class, especially among education and computer science students. This controversy foreshadowed diffficulties that would arise in the development of the instrument to be used in the research component of English 34.

Concurrent with the design of the curriculum and the development of the course in oral communication, a research component became an integral part of the overall project. It was hoped that at the end of a forty-hour, ten-week period of instruction some degree of improvement could be detected in the performance of the students on a specified task. Vide 'aped samples of the English 34 students' speech were collected before the quarter began and then again at the end of the term for the first two quarters the course was offered.

Each prospective student came for an interview. After a three- to four-minute period of general pleasantries, he was asked to select one term from among five which were taken from his major academic field and to explain the term or concept to the interviewer. The student was to role-play. He was asked to think of himself as a teaching assistant and the interviewer as an undergraduate student who was having difficulty understanding the concept. He had five minutes to explain the concept without using visual aids such as a blackboard or paper and pencil. He had to rely on his oral skills alone to communicate. The interviewer, a female native speaker of English, was the same person for all subjects, and identical directions (Appendix A) were given to each subject. The results of rater evaluations of the pre- and post-course interviews are reported elsewhere (Hinofotis, Bailey, and Stern, 1979).

Due to the research component of the English 34 project, the instrument further evolved as an evaluative tool for rating the videotapes of student performance. The ultimate goal is to use the instrument for screening prospective foreign teaching assistants and other students who are interested in taking the course to determine if they need the course and whether they are, in fact, ready for it in terms of their language proficiency. The instrument that was used for evaluating the pre-course and post-course videotapes was developed and revised as part of the pilot study reported here.

The purpose of the pilot study was two fold. First, it provided an avenue for refining the instrument; and second, it allowed us to establish intra- and inter-rater reliability with the instrument. In the pilot study six raters and the three researchers viewed ten videotapes of subjects performing the task described above and evaluated the subjects' performance. The raters were deliberately not trained because we were interested in obtaining unbiased reactions from them

regarding the features of communication that they felt most influenced their ratings of the subjects. Furthermore, we had no predetermined notions about what rating a given subject should receive. In the truest sense, the pilot study was exploratory.

The six raters (three male, three female) who evaluated the videotapes were all trained in the field of teaching English as a second language (TESL), but the amount of actual teaching experience they had had varied from years of experience to very little. The three researchers, who also evaluated the subjects' performance in the pilot study, were experienced ESL teachers and have worked together very closely in the development and implementation of English 34. The effect of their common frame of reference with regard to oral communication is an issue that will be discussed below. Throughout this paper the three researchers are designated as raters 7 through 9 while the six raters unfamiliar with the project are referred to as raters 1 through 6.

The initial draft of the instrument was open-ended and was used by the raters for the first viewing. At the end of each subject's explanation, the raters were asked to indicate their overall impression of the subject's performance by marking the appropriate box on a Likert Scale which had a spread of 1 through 9. (The following verbal descriptions appeared above the numbers: 1, poor; 3, fair; 5, average; 7, good; and 9, excellent.) Next, the raters were asked to provide notes and comments in response to the question, "On what basis did you make this judgment?" Finally, the raters were asked if the subject should be a teaching assistant.

The raters' notes from the first viewing were compiled in an attempt to determine which factors were influencing their evaluation of each subject. On the basis of that information and what we had learned from the teaching-tool stage of the instrument, a draft with performance categories was developed for use during the second pilot study viewing. This draft included three main performance categories — Language Proficiency, Delivery, and Communication of Information — and twelve specific subcategories of performance. Verbal descriptors were written for each of the twelve subcategories. These verbal descriptors and the form of the instrument used for the second viewing are given in Appendix B.

Between the first and second viewings some changes were made in the instrument. However, the overall impression scale was retained so that intra-rater reliability could be established. The question asking whether the subject should be a teaching assistant was revised to focus solely on language ability and communication skills. For the first viewing of the pilot study, the TA question merely asked whether or not the subject should be a teaching assistant at UCLA. Raters found this question difficult to answer because it could refer to the subject's English proficiency, his overall knowledge of his subject, his attitude toward the "student," his willingness to impart information, or all of these areas. The raters pointed out that some of the subjects showed excellent mastery

of their fields, but their problems in English would make it difficult for their students to understand them. On the other hand, some of the subjects were near-native in their English, but were perceived by the raters as being potentially poor teachers because of their apparent attitudes toward the "student." Because of these comments, the question was reworded to ask if the subject's English was good enough for him to be a teaching assistant in his major department at UCLA.

The same nine people viewed the same ten subjects again approximately one month after the initial viewing. For the second viewing they used the newly evolved instrument with performance categories. The subjects had been randomly ordered and numbered 1 through 10 on the videotape for the first viewing. For the second viewing, the raters watched subjects 6 through 10 and then 1 through 5, in order to counteract any ordering effect on the scores. After the second viewing of the tape, additional feedback was elicited from the raters regarding the changes in the instrument and the evaluation process in general. We were especially interested in comments and suggestions about the performance categories on the latest draft of the instrument. The information obtained from the raters and from the data analyses was used to further revise the instrument. These revisions are discussed below.

Data analysis

One of the purposes of this study was to pilot the rating instrument described above. To that end the raters evaluated the videotape samples a second time. Following the second viewing, their responses to the overall impression question, the performance categories, and the TA question were analyzed. In the discussion that follows, each of these three areas of concern is covered in turn.

Global ratings.

In compiling the data obtained from the raters' overall impressions, we were first concerned with establishing intra-rater reliability, an index of each rater's consistency in judging the same performance on different occasions. Using the global scores, a Pearson product moment correlation coefficient was obtained for each rater across the first and second viewing. Table 1 summarizes the results.

For the majority of raters, high correlations were obtained, indicating that most of the raters were consistent in the overall ratings they assigned to the same subject's performance of the task on two different viewings. In fact, considering that the raters were not trained and were given no guidelines for evaluating the subjects, the correlations were impressive. For the combined viewings,

TABLE 1
Intra-rater Reliability Coefficients and Standard Deviations on Overall Ratings for Two Viewings

Rater	Viewing 1 SD	Viewing 2 SC	r
1	2.38	2.17	.96**
2	2.63	2.42	.86**
3	2.51	2.00	.96**
4	2.36	1.14	.71**
5	1.79	1.76	.84**
6	1.78	1.51	.66*
7	2.21	2.00	.78**
8	2.06	2.32	.89**
9	1.60	1.52	.92**

*P <.05
**P <.01

an average intra-rater reliability coefficent of .87 was calculated by using the Fisher Z transformation procedure (Guilford, 1973: 145-146). The overall impression scores for all ten subjects on the nine-point scale were also used to compute the summary statistics given in Table 2.

TABLE 2
Means and Standard Deviations for Nine Raters for Two Viewings

	Viewing 1			Viewing 2		
Raters	$\bar{X}$	Range	S.D.	$\bar{X}$	Range	S.D.
1	5.1	(3-8)	2.4	5.6	(3-8)	2.2
2	5.6	(1-9)	2.6	5.5	(2-9)	2.4
3	6.1	(2-9)	2.5	5.3	(3-8)	2.0
4	6.0	(2-9)	2.4	5.8	(3-7)	1.1
5	5.1	(2-8)	1.8	6.0	(3-8)	1.8
6	6.5	(3-9)	1.8	6.6	(5-9)	1.5
7	4.7	(1-8)	2.2	4.7	(2-9)	2.0
8	5.3	(2-8)	2.1	5.6	(2-9)	2.3
9	5.9	(3-8)	1.6	6.1	(3-8)	1.5

The means and ranges in Table 2 reflect variation in performance perceived by each rater among the subjects. The standard deviations confirm the degree to which individual raters perceived differences among the ten subjects. Given the limited spread of the nine-point rating scale, standard deviations of 2.0 or above may indicate the wide range of oral proficiency of the subjects.

The inter-rater reliability coefficients were computed for both the first and the second viewings across three different combinations of the data: (1) the ratings given by all nine raters, (2) those of the six raters alone, and (3) those of the three researchers. In this case, inter-rater reliability indicates the extent of

agreement among the raters' assessment of the subjects' performance. Table 3 reports the reliability coefficents for all nine raters (1-9), the six raters (1-6), and the three researchers (7-9).

TABLE 3
Inter-rater Reliability Coefficients of Raters' Overall Impressions for Two Viewings

Raters	Viewing 1	Viewing 2
1-9	.89**	.90**
1-6	.78**	.81**
7-9	.95**	.88**

**P <.01

For the first viewing, the reliability coefficient of .95 for the three researchers is extremely high, indicating substantial agreement about the overall speaking ability of the subjects. For the six raters alone, however, the coefficient of .78 is less impressive. The lower coefficient indicates that the six raters evaluated the relative oral abilities of the subjects quite differently. The .89 coefficient for the nine raters reflects inflation caused by the .95 coefficient of the researchers.

As mentioned above, the nine raters in this pilot study viewed the same ten videotapes twice. The order of presentation was altered and a month elapsed between the first and second viewings. Table 4 gives the mean scores representing the nine raters' overall impressions of the same performance by each subject on these two viewing occasions. It also gives the rank ordering of the mean scores.

TABLE 4
Mean Scores and Rank Orderings of the Overall Impressions of Nine Raters for Two Viewings

Viewing 1			Viewing 2		
Rank Order	$\bar{X}$	Subject	Rank Order	$\bar{X}$	Subject
1	8.22	(9)	1	8.11	(9)
2	7.00	(8)	2	7.22	(8)
3	6.33	(1)	3	6.00	(1)
4	6.11	(7)	3	6.00	(7)
5	6.00	(10)	5	5.89	(5)
6	5.56	(3)	6	5.78	(3)
7	5.33	(5)	6	5.78	(2)
8	5.11	(2)	8	5.11	(10)
9	3.56	(4)	9	3.56	(4)
10	2.56	(6)	10	3.44	(6)

The varied mean scores show that the raters did in fact perceive differences among the performances of the ten subjects. The similar rank orderings of the

mean scores for the first and second viewings reveal the consistency with which the subjects' overall English proficiency was evaluated by the nine raters.

Because the videotapes evaluated by the nine raters on the first and second viewings were identical except for ordering, one would predict no significant difference between the mean scores for each subject across the two viewing occasions. The results on an analysis of variance reported in Table 5 support this prediction.

TABLE 5
ANOVA Source Table for Overall Impressions
of Ten Subjects by Nine Raters on Two Different Viewing Occasions

Source	*SS*	*df*	*MS*	*F*
Subjects	361.25	9	40.11	19... **
Raters	40.98	8	5.12	2.47*
Occasions	.45	1	.45	.22
Raters × Occasions	9.00	8	1.13	.54
Residual	317.83	153	2.08	

*P< .05
**P< .01

There were no significant mean differences for subjects across the two viewing occasions. It is interesting to note, however, that there was a significant difference in means among raters in their evaluations of the subjects in spite of the fact that the inter-rater reliability coefficients reported above were generally high.

Regression analysis of discrete variables.

Comments made by several of the raters indicated that it was very difficult to evaluate the subjects on all of the performance categories on the rating sheet, even though all of the categories had been mentioned frequently in the open-ended comments by the same raters on the first viewing. This has led us to consider simplifying the instrument by eliminating categories that provide the least information about the students' overall oral proficiency.

To help determine which categories could be eliminated without a significant loss of information, stepwise regression analyses were run on the data. In the first series of analyses, we wanted to see what combination of subcategories best predicted the ratings on the three major categories — *Language Proficiency*, *Delivery*, and *Communication of Information*. Tables 6 through 8 report the results.

As Table 6 indicates, *Grammar* alone accounts for 76 percent of the variance in the larger *Language Proficiency* category. The addition of *Flow of*

TABLE 6
Statistics for the Regression of *Language Proficiency* on Subcategories

Variable	Multiple R	R^2	Simple R	B	F	Overall F
Grammar	.87	.76	.87	.35	32.09**	196.31**
Flow of Speech	.92	.85	.73	.18	16.63**	
Pronunciation	.94	.88	.86	.27	22.27**	
Vocabulary	.95	.90	.84	.29	18.95**	

**P<.01

Speech to the regression increases the predictability of the *Language Proficiency* rating to .85. The addition of the remaining two variables, *Pronunciation* and *Vocabulary*, increases the amount of variance accounted for in *Language Proficiency* to .88 and .90 respectively. The F ratio associated with each additional variable is significant ($p<.01$), indicating that the combination of the four variables better predicts the *Language Proficiency* rating than a single variable or combination of fewer than four.

TABLE 7
Statistics for the Regression of *Delivery* on Subcategories

Variable	Mutliple R	R^2	Simple R	B	F	Overall F
Enthusiasm	.83	.68	.83	.39	45.60**	97.15**
Eye Contact	.88	.78	.63	.17	15.10**	
Confidence in Manner	.90	.82	.81	.24	11.53**	
Other Nonverbal Aspects	.91	.82	.71	.10	2.16	

**p<.01

Table 7 provides the results of the regression of four variables on the major category of *Delivery*. The subcategory *Enthusiasm* accounts for the largest percent of variance, 68 percent, with *Eye Contact* and *Confidence in Manner* significantly increasing the predictability to 78 and 82 percent respectively. The last variable entered, *Other Nonverbal Aspects*, provided no significant addition to the accounted variation in delivery. It appears that this variable may not be a crucial element in the evaluation of *Delivery* (at least not for the nine raters in the pilot study). However, because the subjects in the study were sitting for the duration of the interview and were accordingly restricted in movement, we are reluctant to eliminate this subcategory until further research is completed.

The combination of the variables reported in Table 8 accounts for 94 percent of the variance in the major category *Communication of Information*. The subcategory *Development of Explanation* alone accounts for 86 percent, with *Ability to Relate to Students*, *Clarity of Expression*, and *Use of Supporting Evidence* increasing the predictability to .91, .93, and .94 respectively. Each addi-

TABLE 8
Statistics for the Regression of
Communication of Information on Subcategories

Variable	Multiple R	R^2	Simple R	B	F	Overall F
Development of Explanation	.93	.86	.93	.35	38.65**	341.71**
Ability to Relate to Students	.96	.91	.82	.22	32.80**	
Clarity of Expression	.97	.93	.91	.22	16.24**	
Use of Supporting Evidence	.97	.94	.90	.18	10.63**	

**p<.01

tional variable has provided a significant increment in the known variation of the major category. *Communication of Information*.

The results of this series of regression analyses indicate that, with the exception of *Other Nonverbal Aspects*, all of the subcategories identified on the rating instrument contribute significantly to the evaluation of subjects in the major categories. The next step involved the regression of the three major categories on the overall rating. The results are reported in Table 9.

TABLE 9
Statistics for the Regression of Major Categories on Overall Rating

Variable	Multiple R	R^2	Simple R	B	F	Overall F
Communication of Information	.88	.77	.88	.46	48.94**	216.31**
Language Proficiency	.94	.87	.80	.37	68.44**	
Delivery	.94	.88	.80	.19	6.31*	

*p<.05
**p<.01

The category *Communication of Information* accounts for the largest single amount of variance (.77) in the overall ratings. The addition of *Language Proficiency* and *Delivery* increases the predictability of the overall ratings to .87 and .88 respectively. Since the F ratio associated with each additional variable is significant, it can be assumed that all three categories were contributing factors in the evaluation of each subject's performance.

Table 10 reports the regression of the subcategories on the overall rating.

In this analysis the ten variables entered in the regression accounted for 88 percent of the variance in the overall ratings, just as the three main category variables did in the previous analysis. Since the main categories taken as a group and the subcategories taken as a group each account for the same substantial amount of variance in the overall score, the question arises as to whether both

TABLE 10
Statistics for the Regression of Subcategories on Overall Rating

Variable	Multiple R	R^2	Simple R	B	F	Overall F
Clarity of Expression (COI)	.85	.73	.85	.17	3.54	55.46**
Grammar (LP)	.89	.80	.69	.19	8.45**	
Confidence in Manner (D)	.92	.85	.73	.18	5.28*	
Development of Explanation (COI)	.92	.86	.82	.13	2.66	
Flow of Speech (LP)	.93	.86	.75	.15	6.22*	
Ability to Relate to Student (COI)	.93	.87	.72	.11	3.35	
Other Non-verbal Aspects (D)	.93	.87	.54	-.96	1.97	
Use of Supporting Evidence (COI)	.94	.87	.76	.99	1.28	
Pronunciation (LP)	.94	.88	.64	.70	1.04	
Eye Contact (D)	.94	.88	.53	-.14	.08	

COI = Communication of Information, LP = Language Proficiency, D = Delivery
*p<.05
**p<.01

types of ratings are needed. Indeed, the overall rating alone seems to provide as much information about the subject's communicative ability in a global sense as the performance categories. Decisions regarding simplification of the instrument must be based on the kind of information that is needed about the student's communicative ability.

It should be noted that all of the regression analyses reported in this paper reflect the reactions of these nine raters towards the ten subjects on videotape. Until further research has been completed with the instrument it is premature to generalize these results beyond the present study.

Analysis of TA question responses.

The final section of the data analysis in this study deals with the raters' responses to the question of whether each subject should be a teaching assistant. The mean scores on the overall impression question were tabulated with the raters' yes/no responses on the TA question. The results are summarized in Table 11. In spite of individual differences, these means provide a first step towards establishing acceptable levels of English proficiency among potential TA's.

In the second viewing, every rater changed his opinion at least once regarding the TA question. Altogether 28 percent of the responses to the TA question changed from the first to the second viewing. Of the total responses, 18 percent

TABLE 11
Mean Overall Scores and Yes/No Responses to the TA Question

Response to TA Question	$\overline{X}$ on Viewing 1	$\overline{X}$ on Viewing 2
Yes	6.50	6.39
No	3.62	3.48

changed from *No* to *Yes*, while only 1 percent of the total responses changed from *Yes* to *No*. Thus, the trend was for the raters to become less critical of the subjects' proficiency with respect to their potential as TA's.

In order to determine the degree of correspondence between the overall impression scores and the yes/no answers on the TA question, a point-biserial correlation coefficient (r_{pb}) was computed for the first and second viewings. Correlations of .79 for the first viewing and .70 for the second viewing were obtained by using the Fisher Z transformation procedure (Guilford, 1973: 145-146). These figures may be interpreted in the same way inter-rater reliability coefficients are interpreted. The coefficient indicates the extent to which the score on a continuous variable (the nine-point global scale) correlates with the "score" on the dichotomous variable (the yes/no answer on the TA question).

Correlation coefficients of .79 and .70 are positive but not particularly high. However, when the point-biserial correlation coefficient was calculated for each individual rater across all subjects, considerable variation among the raters was found, as shown in Table 12.

TABLE 12
Point-biserial Correlation Coefficients (r_{pb}) for Acceptability Decisions vs. Overall Scores

Rater	First Viewing	Second Viewing
1	—*	.73
2	.72	.71
3	.58	.89
4	.91	.87
5	.76	.60
6	.58	.37
7	.86	.32
8	.88	.70
9	.79	.71

*In the first viewing, no r_{pb} could be computed for Rater 1 because he awarded "yes" answers to all the subjects on the TA question.

The point-biserial correlation coefficients reported in Table 12 may be read as measures of the systematicity (i.e., the intra-rater reliability) of each rater's overall scores and yes/no responses to the TA question.

The point-biserial correlations reveal a potential problem in the rating process. An area of consideration which comes into play in most research involving raters is the fatigue factor. Fatigue probably affected some of the raters during the second viewing, which took place late in the day. For example, the point-biserial correlation for Rater 7 dropped dramatically from the first viewing to the second. The rater attributes this to fatigue. This problem was eliminated in subsequent rating sessions.

Discussion

This paper has reported on the development of a rating instrument for measuring the oral English proficiency of nonnative applicants for teaching assistantships. Both the subjective feedback of the raters and the information gained in the data analysis have been used in revising the instrument.

The nine-point rating scale on the overall impression question has been retained and now appears twice on the latest version of the instrument, as "Initial Overall Impression" and "Final Overall Impression." In the future this format will be used in an attempt to determine how evaluation on the performance categories influences the overall ratings.

The descriptions for the twelve subcategories were revised based on comments elicited from the raters following the second viewing of the pilot tapes. For example, the subcategory *Enthusiasm* (which had emerged as a very important area in the raters' open-ended comments from the first viewing) presented a number of problems. Some of the raters felt that appropriate degrees of enthusiasm might vary from one discipline to another, as well as from one culture to another. In addition, this topic seems to be one area in which the interviewer's involvement potentially influenced the subject's performance. The category was originally called *Enthusiasm* and the descriptor read "apparent degree of interest in sharing knowledge with the 'student'." This wording was seen as being somewhat nebulous. It was revised to read, "Apparent degree of animation and enthusiasm, as reflected in part by voice quality; may include use of humor." The category was retitled *Presence* in hopes that this term would convey those aspects of personality which seemed to provoke an affective response among the raters.

A major area of interest in the overall project is the use of this instrument for screening foreign students who are applying for teaching assistantships. For this reason, the TA question is extremely important. In the pilot study several raters pointed out that teaching assistants have different responsibilities, depending upon their major departments, and these differing responsibilities may demand different levels of English proficiency. UCLA's TA Manual supports this notion of differing responsibilities by classifying TA's into three major roles: instructing one's own class, leading a discussion section, and conducting a labo-

ratory section. Considering this distinction, the TA question was revised as follows:

Is this subject's English good enough for him to be a teaching assistant in his major department at UCLA in the following capacties? (Please circle *yes* or *no*.)

A. Lecturing in English	Yes	No
B. Leading a discussion section	Yes	No
C. Conducting a lab section	Yes	No

The instrument incorporating the revisions discussed above (Appendix C) has been used in a follow-up study (Hinofotis, Bailey, and Stern, 1979). The results to date are encouraging; however, continued revisions are planned pending subsequent data analyses.

This pilot study has suggested a number of areas for further research with the instrument. Since undergraduate students comprise the population most affected by foreign TA's, we plan to have undergraduates from a variety of disciplines evaluate the subjects already on videotape. We would also like to have the videotape data evaluated by faculty members who are involved in TA selection in various departments. Also, a question remains as to whether Teaching Assistants in different disciplines need the same language proficiency and communicative skills. A natural step in providing baseline data would be to evaluate the performance of native-speaking TA's using the same criteria. Finally, we hope to use the instrument in live observations in the classrooms of foreign TA's.

However, an evaluation instrument is only one facet of a performance test of oral proficiency. The data collection process must also be examined. An issue of concern is the extent to which a role-play task, such as the one used in this study, can predict a nonnative applicant's potential as a teacher in his major area. It may be that we have tapped some role-play ability as well as oral proficiency. The relative breadth and complexity of the terms explained by the subjects is yet another unexplored variable. The technical aspects of data collection (e.g., the camera angle in videotaping) introduce additional methodological questions. Furthermore, the role of the interviewer (i.e., the mock "student" in these data) seems to influence the raters in judging the subject's performance. Finally, the question of measuring communicative versus linguistic competence of prospective foreign TA's dictates the need for a thorough job analysis by disciplines. All of these issues merit futher examination.

This study has been conducted to pilot a rating instrument for measuring oral English proficiency in a simulated teaching situation. It is our hope that further refinements of the instrument will provide a measurement component which can be used in a performance test of oral proficiency for screening foreign applicants for teaching assistantships.

REFERENCES

Guilford, J. P. and B. Fruchter. 1973. *Fundamental statistics in psychology and education*, 5th ed. New York: McGraw-Hill.

Hinofotis, F. B. and K. M. Bailey. 1978. Course development: oral communication for advanced university ESL students. In J. Povey, ed., *UCLA workpapers in teaching English as a second language XII*. 7-20.

Hinofotis, F. B., K. M. Bailey, and S. L. Stern. 1978. A progress report on English 34: oral communication for foreign students. Unpublished manuscript. Los Angeles: Department of English (ESL Section), University of California.

_____. 1979. Assessing improvement in oral communication: raters' perceptions of change. In J. Povey, ed., *UCLA workpapers in teaching English as a second language XIII*.

Jones, R. 1979. Performance testing of second language proficiency. In E. J. Brière and F. B. Hinofotis, eds., *Concepts in language testing: some recent studies*. Washington, D.C.: Teachers of English to Speakers of Other Languages. 50-57.

APPENDIX A
Instructions to Subjects

Here are five terms related to your academic field. Choose one you would feel comfortable explaining. (The students were allowed to reject all five terms and choose from five others if they wished. This process continued until each student found a vocabulary item with which he/she was familiar. We were flexible in this matter because we wanted to measure the students' abilities to explain familiar material, rather than test their knowledge of the subject matter.)

Imagine that you are the teaching assistant for an introductory _____ course, and that I am a student in the class. I missed a lecture and I have come to you for help before an examination. I don't know this term, which I came across in my reading, and I think it will be on the test. You have five minutes to explain this term to me in any way you can without writing or drawing anything. You can take some time to think about what you'll say. Do you have any questions?

APPENDIX B
Descriptors and Rating Instrument Used in Pilot Study

Descriptors

During the second viewing of the pilot videotapes, you will be asked to rate the subjects in several specific categories. These topics and the areas they cover are listed below. You may refer to this sheet during the rating process if you wish. Please make any suggesions that would help us clarify these categories on the attached rating form.

1. **Vocabulary**, including semantically appropriate word choice, control of idiomatic English, and subject-specific vocabulary.
2. **Grammar**, including the morphology and syntax of English.
3. **Pronunciation**, including vowel and consonant sounds, syllable stress, and intonation patterns.

4. **Flow of Speech**: smoothness of expression, including rate and ease of speech.
5. **Eye Contact**: looking at the "student" during the explanation.
6. **Other Nonverbal Aspects**, including gestures, facial expressions, posture, freedom from distracting behaviors, etc.
7. **Confidence in Manner**: apparent degree of comfort or nervousness in conveying the information.
8. **Enthusiasm**: apparent degree of interest in sharing knowledge with the "student."
9. **Development of Explanation**: degree to which ideas are coherent, logically ordered, and complete.
10. **Use of Supporting Evidence**, including spontaneous use of example, detail, illustration, analogy, and definition.
11. **Clarity of Expression**, including use of synonyms, paraphrasing, transitions, level of diction, and precise word choice.
12. **Ability to Relate to "Student,"** including apparent attitude, degree of flexibility in responding to questions, and monitoring of student's understanding.

English 34 Rating Instrument

Subjects's number: _____ Rater's number: _____

Term being defined: _____ Date: _____

Directions: You will see a series of videotaped interviews in which each subject explains a term from his/her academic field. As the tape is playing, you may make notes about the subject's performance in the space below, in order to help you arrive at an overall rating. When the tape ends, please give your overall impression of the subject's performance of the task by marking the appropriate box under "Overall Impression." Then answer the question below. After you have done this, please turn over the page and fill out the checklist.

Overall Impression

Poor		Fair		Average		Good		Excellent
☐	☐	☐	☐	☐	☐	☐	☐	☐
1	2	3	4	5	6	7	8	9

	Yes	No
Is this subject's English good enough for him/her to be a T.A.?	☐	☐

Optional comments:

	POOR 1	2	FAIR 3	4	AVERAGE 5	6	GOOD 7	8	EXCELLENT 9
LANGUAGE PROFICIENCY									
1. Vocabulary									
2. Grammar									
3. Pronunciation									
4. Flow of speech									
DELIVERY									
5. Eye contact									
6. Other nonverbal aspects									
7. Confidence in manner									
8. Enthusiasm									
COMMUNICATION OF INFORMATION									
9. Development of explanation									
10. Use of supporting evidence									
11. Clarity of expression									
12. Ability to relate to "student"									

APPENDIX C
Revised Descriptors and Rating Instrument

Descriptors

In viewing the videotapes, you will be asked to rate the subjects in three general categories and twelve specific categories. These topics and the areas they cover are listed below. You may refer to this sheet during the rating process if you wish.

A. LANGUAGE PROFICIENCY

1. **Vocabulary**, including semantically appropriate word choice, control of idiomatic English, and subject-specific vocabulary.
2. **Grammar**, including the morphology and syntax of English.
3. **Pronunciation**, including vowel and consonant sounds, syllable stress, and intonation patterns.
4. **Flow of Speech**: smoothness of expression, including rate and ease of speech.

B. DELIVERY

5. **Eye Contact**: looking at the "student" during the explanation.
6. **Other Nonverbal Aspects**, including gestures, facial expressions, posture, freedom from distracting behaviors, etc.
7. **Confidence in Manner**: apparent degree of comfort or nervousness in conveying information.
8. **Presence**: apparent degree of animation and enthusiasm, as reflected in part by voice quality; may include humor.

C. COMMUNICATION OF INFORMATION

9. **Development of Explanation**: degree to which ideas are coherent, logically ordered, and complete.
10. **Use of Supporting Evidence**, including spontaneous use of example, detail, illustration, analogy, and/or definition.
11. **Clarity of Expression**, including use of synonyms, paraphrasing, and appropriate transitions to explain the term; general style.
12. **Ability to Relate to "Student,"** including apparent willingness to share information, flexibility in responding to questions, and monitoring of "student's" understanding.

Oral Communication Rating Instrument

Subject #_____ Term _____ Date _____ Rater # _____

Directions: You will see a series of videotaped interviews in which each subject explains a term from his/her academic field. As the tape is playing, make notes about the subject's performance of the task in the space below. When the tape ends, please give your *initial overall impression* of the subject's performance by circling the appropriate number under Roman numeral I. After you have done this, please turn over the page and complete Roman numerals II and III in sequence.

I. Initial Overall Impression

Please circle only one number:

1 (Poor) 2 3 4 5 6 7 8 9 (Excellent)

II. Oral Communication Performance Categories

Directions: Rate this subject on each of the following fifteen categories. Please circle *only one* number for each category.

	(Poor)								(Excellent)
LANGUAGE PROFICIENCY	1	2	3	4	5	6	7	8	9
1. Vocabulary	1	2	3	4	5	6	7	8	9
2. Grammar	1	2	3	4	5	6	7	8	9
3. Pronunciation	1	2	3	4	5	6	7	8	9
4. Flow of speech	1	2	3	4	5	6	7	8	9
DELIVERY	1	2	3	4	5	6	7	8	9
5. Eye contact	1	2	3	4	5	6	7	8	9
6. Other nonverbal aspects	1	2	3	4	5	6	7	8	9
7. Confidence in manner	1	2	3	4	5	6	7	8	9
8. Presence	1	2	3	4	5	6	7	8	9
COMMUNICATION OF INFORMATION	1	2	3	4	5	6	7	8	9
9. Development of explanation	1	2	3	4	5	6	7	8	9
10. Use of supporting evidence	1	2	3	4	5	6	7	8	9
11. Clarity of expression	1	2	3	4	5	6	7	8	9
12. Ability to relate to "student"	1	2	3	4	5	6	7	8	9

III. Final Overall Impression 1 2 3 4 5 6 7 8 9

Is this subject's English good enough for him to be a teaching assistant in his major department at UCLA in the following capacities? (Please circle *yes* or *no*.)

A. Lecturing in English	Yes	No
B. Leading a discussion section	Yes	No
C. Conducting a lab section	Yes	No

Optional Comments:

*Measurements of Reliability and Validity of Two Picture-Description Tests of Oral Communication**

Adrian S. Palmer
University of Utah

Abstract. Since Upshur's (1969) original paper describing a picture-description test of oral communication ability, four empirical studies have been completed in which variants of the test have been used. From these studies considerable new information on the tests' reliability and validity has become available. Indications are that the reliability is somewhat less than originally estimated and that concurrent validity with the oral interview is disturbingly low. A feature analysis of the speech behavior required by the tests indicates a number of abnormalities which could account for the tests' low validity. The implication is that if controlled tests of communication are needed, an effort should be made to minimize the effect of controls on the naturalness of the speech behavior.

Introduction

Ten years ago, John Upshur presented a paper entitled "Measurement of oral communication" (Upshur, 1969) in which he described a particular method of testing oral communication involving timed picture-description tasks. Since then, four studies have been completed in which this method of testing has been used. In the first of these studies two variants of Upshur's test were analyzed for reliability and factoral structure; in the subsequent three studies these tests were used in research in second language acquisition. This paper reviews the published findings, presents some new data on test reliability and validity, analyzes the test method, and offers some general conclusions about the usefulness of the tests.

*The author wishes to thank George A. Trosper for his comments.

The Research

Description of the tests

PROTEST. PROTEST is a test of oral production. It is an adaptation of Upshur's picture description test, a test in which the testee is shown four similar pictures and told to describe one of them in a single sentence. His response is recorded and later played to a native speaker auditor. The auditor decides which picture he thinks has been described, and the response is scored either correct or incorrect depending on the match between the auditor's judgment and the testee's intent. In addition, a record is kept of the length of time required for the testee to complete his description.

PROTEST differs from Upshur's test in that, instead of recording his description, the testee describes his picture directly to the examiner. If the testee either fails to provide enough information for the examiner to identify one picture from the four, or if he provides information leading to the examiner's incorrectly identifying the described picture, the examiner provides feedback to the testee which requires that he continue his description until the correct picture can be identified. Thus, inaccuracies in the propositional content of the testee's description are automatically converted into increased time to complete the task. As a result, only one type of "score" is recorded: the amount of time necessary for the testee to describe the designated picture so that the examiner can identify it correctly.

COMTEST. COMTEST is a test of two-way oral communication using the same four-picture cards. The basic task is for the testee to ask an examiner a series of questions (yes/no or either/or) to determine which of the four pictures the examiner has in mind. The testee continues asking questions until he has correctly identified the "key" picture. His score is the amount of time required to complete this task.

The actual procedures for administering this test are somewhat more complicated, the complexity resulting from the need to eliminate chance as a factor in performance. If, for example, a testee were to start by asking about the particular picture that the examiner had in mind, he would be able to identify this picture rather quickly — perhaps with only one question. If, however, he were to ask about the correct picture last, up to four questions — and a considerably longer time — would be required. As a result, a testing procedure was developed which would allow the testee to identify the "correct" picture only after he had asked three informative questions, questions sufficiently explicit to allow the examiner to figure out which particular picture(s) the testee was trying to accept/reject with his question.

The key to the procedure is for the examiner actually *not* to have any particular picture in mind. Instead, he follows a procedure for answering informative questions which insures that the testee will not have sufficient information

to identify unconditionally any one of the pictures as correct until he has asked three informative questions and understood their answers. Once the testee has done so, the examiner answers the testee's final question in such a way that the testee can eliminate all but one picture from consideration — thereby allowing him to identify the "correct" picture.

Reliability studies

Study #1. The reliability of PROTEST and COMTEST were first measured in a 1972 study (Palmer, 1972). Both tests were administered to 33 nonnative speakers and five native speakers at the English Language Institute, University of Michigan. Also administered were the Michigan Test Battery (including composition, listening comprehension, and objective tests of grammar, vocabulary, and reading comprehension) and an experimental listening comprehension test. The reliabilities of PROTEST and COMTEST were estimated by computing multiple correlation coefficients, with each of the two experimental communication tests as dependent variables.

Multiple R's for PROTEST and COMTEST (ten-item tests) were .82 for PROTEST and .80 for COMTEST. These values were taken as lower-bound reliability estimates, the assumption being that whatever portions of the variances on PROTEST and COMTEST were predictable must be reliable. The Spearman-Brown prophecy formula was then used to estimate lower-bound reliability for double test length (twenty-item tests) and determined to be .90 for PROTEST and .89 for COMTEST.

Two factors, however, may have contributed to inflated multiple R's in this study. One factor is sampling fluctuation, which can be eliminated with a cross-validation study. McNemar (1969: 208) suggests using the regression equation based on the first sample to calculate predicted values for the subjects in a second sample. Then, by correlating these predicted values with the obtained values of the second sample individuals, one can determine the worth of the initial multiple regression equation (and multiple R). However, since no cross-validation study was performed, it is impossible to know the extent to which the obtained multiple R's were inflated due to sampling fluctuation.

Multiple R's may also be inflated if the number of predictors is fairly large relative to the number of subjects. Guilford (1965: 40) provides a formula for calculating shrinkage due to this factor: $R^2 = 1 - (1-R^2)(N-1/N-m)$. When this formula is applied to the data in the 1972 study, the obtained corrected values of R are somewhat smaller, and when the Spearman-Brown prophecy formula is applied to these corrected values to estimate reliablity for double test length, lower reliability estimates are obtained. Uncorrected and partially corrected reliability estimates for PROTEST and COMTEST are given in Table 1.

While these predicted values of reliability for twenty-item PROTEST and COMTEST are somewhat reduced, they are undoubtedly still inflated due to

TABLE 1
Uncorrected and Partially Corrected
Reliability Estimates for PROTEST and COMTEST

Measure	Uncorrected Multiple R	Partially Corrected Multiple R	Partially Corrected Reliability Estimates For 20-item Tests
PROTEST (10-item)	.82	.76	.86
COMTEST (10-item)	.80	.73	.84

sampling fluctuation. The extent of this overestimation can be seen in reliability figures obtained in a second study.

Study #2. Twenty-item versions of PROTEST and COMTEST were used as part of an experiment in teaching for acquisition in a foreign language environment (Palmer, 1978a). The two tests were administered to 60 second-year students in a Thai university. Alternate items in COMTEST were scored separately, and Rulon's statistic (Guilford, 1965: 445) was used to compute the standard error of measurement directly from differences between scores of individuals on odd and even pools of items from the same test. The reliability of COMTEST obtained by this method was .64, a value considerably less than the .89 (uncorrected) and .80 (partially corrected) values obtained in Study #1. Moreover, since the estimated reliabilities for PROTEST and COMTEST in Study #1 were nearly the same, it seems reasonable to assume that the "true" reliability of PROTEST is also on the order of .64, rather than the higher value obtained in Study #1.

Studies #3 and #4. PROTEST and COMTEST were also administered on two more occasions. In Study #3, the tests were used as part of a test battery to measure accuracy, communicativity, and social judgments for two groups of Thai foreign language learners (Upshur and Palmer, 1974; Palmer, 1978a). In this study, the tests were given to 24 Thai housemaids and 24 Thai university students.

In Study #4, the tests were used as part of an experiment in teaching for acquisition in an EFL classroom (Palmer, 1978b). Here, the tests were administered to two groups of 26 subjects. The subjects, first-year engineering students in a Thai university, had been taught English for one semester in two different ways (following a number of years of similar high school instruction).

Intercorrelations between PROTEST and COMTEST in Studies #1-#4 can be used as indirect estimates of their reliabilities for three reasons. First, the test method and the content of the two tests are, for all practical purposes, identical. Both use similar sets of pictures, require similar types of speaking behavior, and use similar scoring procedures. Second, the testees' speech behavior is very similar in both tests (see the features of autonomous communication de-

scribed below). Third, the factoral structures of the two tests (Palmer, 1972) are very similar. Therefore, it seems reasonable to consider PROTEST and COMTEST alternative forms of the same test.

If two tests measure the same thing, and if their reliabilities are the same (as they were found to be, for all practical purposes, in Study #1), the obtained correlation between the tests cannot be greater than the reliability coefficient (McNemar, 1969: 172). Thus, the correlation between PROTEST and COMTEST can be taken as an upper-bound estimate of the reliability of the two tests. The intercorrelations of PROTEST and COMTEST in Studies #1-#4 are given in Table 2.

TABLE 2
Intercorrelations Between PROTEST and COMTEST
in Studies #1 through #4

Study	Intercorrelations Between PROTEST and COMTEST
Study #1: 10-item tests (N = 38)	.76
Study #2: 20-item tests (N = 60)	.62
Study #3: 20-item tests	
Group 1 (N = 24)	.65
Group 2 (N = 24)	.79
Study #4: 20-item tests	
Group 1 (N = 26)	.44
Group 2 (N = 26)	.67

These intercorrelations lead one to conclude that the reliabilities of the tests are closer to the .64 estimate (using Rulon's statistic in Study #3) than to the uncorrected .90 estimate or to the partially corrected .85 estimate (based upon the multiple R correlations in Study #1.)

Concurrent validity study

In Study #2, the 60 subjects were also interviewed. A panel of three native speakers of English talked with each subject for a total of approximately ten minutes. The subjects were rated on the following scales: pronunciation, grammar, fluency, comprehension, confidence, and social status. The ratings on all of the scales were summed across raters to provide a global score for each subject.

The scores on the subparts of the interview correlated very highly (in the .80-.90 range), and the total interview scores correlated fairly well with a dictation test (.70). However, the correlations of the interview total with PROTEST and COMTEST were low, as seen in Table 3.

TABLE 3
Correlations of PROTEST and COMTEST with Oral Interview Scores in Study #2

	COMTEST	INTERVIEW
PROTEST	.62	.45
COMTEST		.34

These low correlations indicate that PROTEST and COMTEST provide a different type of information from that obtained in the oral interview.

Construct Validity of PROTEST and COMTEST

One way of investigating construct validity is to examine the effects of trait and method on test scores. While none of the studies considered in this paper was designed explicity for this purpose, the studies do provide some indication that method variance introduced in the picture-description tests heavily influenced test scores.

If PROTEST, COMTEST, and the interview had all measured the same trait (oral proficiency), one would have expected all three tests to correlate to the extent that their reliabilities permit. However, despite the fact that the reliabilities of the three tests were in the .60-.70 range (from which one would predict intercorrelations of the same magnitude), a different pattern of relationships emerged.

The data in Table 3 indicated that PROTEST and COMTEST correlated much more highly with each other than with the interview. This rather disquieting situation can be explained in several ways. On the one hand, the experimental tests and the interview may have measured either different traits altogether or different components of a complex trait. Or, if the two types of tests did measure the same trait, the variances in scores introduced by the different methods of testing may nevertheless have been sufficiently large to obscure the common trait variance.

A systematic investigation of the relative importance of these two sources of unique variance in scores on the two types of tests would require both detailed models of trait and method and evidence from a complex, large-scale research study, neither of which is available. One can, however, analyze the speech produced in the picture description tests in terms of a feature model of

autonomous communication. If the analysis indicates substantial differences between this behavior and the speech behavior in the oral interview, one might attribute the low intercorrelations of test scores, at least in part, to these differences.

Features of autonomous communication

This model was developed to highlight the differences among manipulative, meaningful, and pseudo-communicative drills, and autonomous communication. Highly derivative, it is based on the fundamental elements of Searle's (1969) speech act theory, a modification of Harvey's (1977) analysis of communication, Paulston's (1970) classification of structural pattern drills, and Rivers' (1969) analysis of pseudo-communication and autonomous communication.

In this model, as in other similar models, production is seen as varying from pure manipulation at one extreme, through noncommunicative (yet meaningful) use, to autonomous communication at the other extreme. Purely manipulative language use is distinguished from meaningful use by the absence of three features, each of which adds an element of meaningfulness, and meaningful use is distinguished from autonomous communication by the absence of four features, each of which adds an element of communicativity. Between the purely manipulative extreme and the fully meaningful, yet noncommunicative, middle ground is an area characterized here as "semi-meaningful" language use. Likewise, between the fully meaningful mid-point and the autonomous communication extreme is an area characterized (traditionally) as "pseudo-communication." The model is given in Figure 1, and the features involved are discussed below as they apply to PROTEST and COMTEST.

FIGURE 1
Features of Communication

Pure Manipulation	*Semi-Meaningful Manipulation*	*Meaningful Manipulation*	*Pseudo-Communication*	*Autonomous Communication*
− propositional content	± propositional content*	+ propositional content	± propositional content**	+ propositional content
− speech-act content	± speech-act content*	+ speech-act content	± speech-act content**	+ speech-act content
− information sequence	± information sequence*	+ information sequence	+ information sequence**	+ information sequence
− uncertainty	− uncertainty	− uncertainty	+ uncertainty	+ uncertainty
− intent	− intent	− intent	± intent	+ intent
− processing	− processing	− processing	± processing	+ processing
− shared reference	± shared reference	± shared reference	± shared reference	+ shared reference

*Either 1 or 2 of the 3 features so marked must be positive in value.
**Either 1, 2, or 3 of the 3 features so marked must be positive in value.

Propositional content. If the speakers are required to pay attention to the surface meaning of the sentences in the exchange, the exchange is [+ propositional content]. The exchanges in PROTEST and COMTEST are clearly meaningful at this level, since the propositional content of each utterance is based upon a picture and must be verified against that picture.

Speech-act content. To the extent that the speakers are required to pay attention to the purposes of each utterance and to process each utterance for its purpose, the exchange is [+ speech-act content]. To insure that this processing takes place, it is important both that there be a potential for a *variety* of speech acts and that the *order* of the speech acts not be completely predictable. PROTEST and COMTEST clearly do not meet this criterion. In PROTEST, all of the testee's utterances are statements, used simply to provide information, and the examiner is also limited to one speech act: expressing satisfaction or dissatisfaction with the testee's information. In COMTEST, the purpose of each of the testee's utterances is to obtain information which will enable him to accept or reject a particular picture or subset of pictures (although this can be done with a variety of sentence types, including yes/no questions or various types of statements which — in the context of this test — get interpreted as questions); and here also the examiner is limited to one variety of speech act, confirmation or negation. There is no place in either test for any other speech acts, such as apologies, greetings, orders, promises, etc. Thus, the exchanges in both tests are [− speech-act content].

Information sequence. "Information sequence" is a term used by Oller and Obrecht (1969) to distinguish two types of exchanges. An information sequence consists of an extended series of utterances, each of which is responsive to the previous one. The other type of exchange consists either of a series of utterances totally unrelated in information content, yet perhaps related in grammatical structure, *or* a series of utterances consisting merely of pairs of related utterances but not of larger units. (The latter alternative for this second type is mine, not Oller and Obrecht's. It is intended, e.g., to appropriately characterize meaningful drills (Paulston, 1970) which involve single question-answer exchanges carried out a number of times between a teacher and a series of students. Such drills are [− information sequence] according to this second condition.) Thus, the exchanges in PROTEST are ideally [− information sequence] since the testee — assuming he provides an adequate description on his first attempt — produces only a single utterance, and whatever he says in the following problem is unrelated to it. Even if the testee's first description is not adequate, the resulting exchange is only marginally [+ information sequence].

In COMTEST, the exchange would appear to meet the criterion for information sequence to a limited extent since the testee must incorporate the information in the examiner's reply when framing the second and third questions. With most testees, however, the information sequence is extremely simple, e.g.: "Is this or this right?" "Neither." "Is this right?" "No." "Is this right?"

"Yes." There is clearly very little richness in the variety of relationships between successive utterances in this type of exchange.

Uncertainty. If there is uncertainty in the exchange, neither participant can predict in advance exactly what the other will say. Basic to all definitions of communication, this criterion must be met for either pseudo-communication or autonomous communication. Clearly PROTEST and COMTEST meet the criterion, since neither the testee nor the examiner knows exactly what the other will say. However, while the tests are [+ uncertainty], they are not fully meaningful because, as indicated above, there is no variation at the speech-act level. Since degree of uncertainty is related to meaningfulness, a lack of meaningfulness at either the propositional-content level or at the speech-act level will reduce the potential for uncertainty. The degree of the testee's uncertainty is particularly minimal; there are only two possibilities for propositional content in the examiner's utterances (satisfaction vs. dissatisfaction in PROTEST, confirmation vs. negation in COMTEST) and the examiner's speech acts are totally predictable.

Intent. If both participants have full control over the decisions (a) whether or not to communicate and (b) what to communicate, the exchange is [+ intent]. In PROTEST and COMTEST, as in most forms of pseudo-communication, the speakers are forced to communicate. While in some more advanced forms of pseudo-communication (such as role plays, etc.) the speakers *may* get so caught up in the activity that they would continue communicating even if they did not have to, such is probably not the case with PROTEST and COMTEST. Thus, both tests are [− intent].

Processing. If the speakers have full control over *all* the language elements used in their production (vocabulary, syntax, and phonology), and if (as listeners) they must pay attention to *all* the elements in the messages they hear, the exchange is [+ processing]. (I am ignoring here the predictability present in natural speech and focusing only on the additional predictability introduced in controlled speech.) Most communicative drills are [− processing] since large portions of the utterances are generally repeated from exchange to exchange. When taking PROTEST and COMTEST, some testees in fact avoid nearly all processing, choosing instead to produce telegraphic utterances containing only one or two key vocabulary items. Moreover testees are exposed to model questions and statements in the instructions to the tests and can avoid most of the processing by simply sticking to these structures. Likewise, once the examiner gets used to a testee's strategy, he can usually predict the structure the testee will use and pay attention only to a key vocabulary word. In a[illegible] the examiner can predict *a priori* that testees' utterances will usually [illegible] questions in COMTEST and declarative statements in PROTEST. [illegible] the tests usually entail only minimal processing.

Shared reference. Communication is richer when the two participants share considerable experience relevant to the topic since each utterance can evoke a

wide range of responses — responses to the many implications of a particular statement or question. Thus, if I were to tell a motorcycle enthusiast that my Ducati road racer has a desmodromic valve train, he could respond in a wide variety of ways: e.g., "I thought spring systems had caught up with desmos," or "I'll bet it's a pain to adjust," or "Where do you get the closing shims?" or "Is it like the old Mercedes system?" On the other hand, considerable experience in boring my friends and colleagues at the office testifies that the same comment to a nonenthusiast leads either to an abrupt change in topic or to a series of very general, polite questions about what a desmodromic valve train is.

In PROTEST and COMTEST, the pictures provide both examiner and testee with a single frame of reference which helps keep the conversation moving until the communication objective is reached. However, in these tests there may be too much shared reference — the effect being to reduce the demands on the testee's linguistic competence.

Comparison of PROTEST and COMTEST with oral interview

The results of this analysis of PROTEST and COMTEST are summarized in Table 4.

TABLE 4
Analysis of the Communicativity
of PROTEST and COMTEST

FEATURE	TEST	
	PROTEST	COMTEST
propositional content	+	+
speech act content	–	–
information sequence	or marginal	marginal
uncertainty	+	+
intent	–	–
processing	marginal	marginal
shared reference	+ (but not rich)	+ (but not rich)

The number of minuses in Table 4 raises the possibility that the "communication" in PROTEST and COMTEST and that in the interview test are rather different, for an analysis of the speech behavior in an interview would yield plusses for all of the features (with the probable exception of "intent").

Reactions to the Test

This analysis of PROTEST and COMTEST helps explain some of the misgivings that various examiners and reviewers have had about the tests. One observation has been that there seem to be rather substantial differences in the amount of processing required in the tests and during autonomous communication. Since very little processing is *necessary* in the tests, some testees who can barely communicate in autonomous situations are able to speak in the "bizarre mode" (Krashen, 1978); that is, they can use conscious rules or L1 structures to initiate production and plug in L2 vocabulary forms as required. This mode of speaking is hardly representative of that used in natural speech, and the examiners have questioned whether this type of performance is worth testing.

The examiners have also commented that reasoning ability seemed to be an important factor in performance. In autonomous communication, the information in each utterance frequently opens up a fairly wide range of possible responses. In the tests, however, the possible responses are limited according to the results of deductive reasoning. Examiners have frequently noted that fast reasoners frequently perform far better than could be predicted from their acquired control of English. On the other hand, slow reasoners able to perform well in autonomous communication frequently seemed to be inappropriately penalized. They often became perplexed by the implications of the utterances rather than by the English language *per se*. Keith Morrow's (1977) observation that deductive reasoning ability might be an unnaturally important element in this test is undoubtedly correct.

"Teachability," a third problem noted by examiners, stems both from the importance of deductive reasoning ability and from the narrowness of the propositional meaning communicated. A few minutes' practice in drawing conclusions about "possibly correct" pictures from various statements and questions about the pictures seemed to produce a great improvement in some testees' performance. In addition, when a testee's performance seemed to be limited by his control of English, it could be improved quickly by teaching the testee the vocabulary of spatial relationships and by instructing him to treat each picture as a collection of shapes and lines rather than as a representation of an object or an event. This teachability problem would appear to limit the usefulness of the test to one-time administrations for research purposes and to preclude its regular use in the evaluation of instruction.

A final comment has been that the scoring method used in PROTEST and COMTEST prevented the examiners from obtaining more than one type of information about the testee's oral proficiency: his ability to transmit information. This contrasts sharply with the scoring flexibility of the oral interview method. In a suitably structured interview, the testee can be evaluated not only on (a) his ability to use the spoken language to transmit information, but also on (b) his control of language elements (linguistic accuracy) and (c) his control of the social rules of language use.

The scoring method used in PROTEST and COMTEST is incapable of providing the latter two types of information. Indeed, some of the best performers on PROTEST and COMTEST were those testees who used a highly simplified telegraphic style to communicate the minimum amount of information required at a very rapid rate. Where this aspect of language control is the only aspect of a testee's oral proficiency that needs to be measured, the four-picture test method may well be adequate. Where other types of information are needed, however, more sophisticated test methods will be required.

Conclusions

Seven years' experience of using picture-description tests and analyzing the results leads to the following conclusions. First, these tests are only moderately reliable, certainly not reliable enough for use in making decisions about individuals. Thus, their use should be restricted to situations where information is needed about large numbers of testees.

Second, the concurrent validity of the tests is low. They fail to correlate well with the oral interview, the most widely accepted type of oral proficiency test. Insight into the reasons for the tests' failure to correlate with tests of more autonomous communication may be found in a feature analysis of the components of autonomous communication and the tests' rather dismal performance when evaluated by this model.

Third, test-wiseness appears to play an important role in performance. As a result, the tests should probably not be used more than one time for a given population.

On the brighter side, the tests have proven quick and easy to administer. They make few demands on the examiner, requiring him to listen only for one thing (propositional content) and to keep track of only one variable (time), and the time and facilities required to train examiners are minimal. Moreover, the attempt to analyze the nature of these tests' limitations has led to yet another use for a model of pseudo-communication.

Finally, although PROTEST and COMTEST have not held up well under statistical or logical analysis, one should not infer that all pseudo-communication tests are, or need be, equally deficient. Where controlled tests of pseudo-communication are needed, effort should be spent in obtaining the desired degree of control while minimizing the effect of method on the naturalness of the speech behavior.

REFERENCES

Guilford, J. 1965. *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Harvey, J. 1977. Talk given at Brigham Young University, Provo, Utah, November 18, 1977. Dittoed.

Krashen, S. 1978. Adult second language acquisition and learning: a review of theory and applications. Paper presented at the Sixth Annual SPEAQ Conference, Quebec, Canada, June 17, 1978.

McNemar, Q. 1969. *Psychological statistics*. New York: John Wiley and Sons.

Morrow, K. E. 1977. *Techniques of evaluation for a notional syllabus*. Reading: Centre for Applied Language Studies, University of Reading. Study commissioned by the Royal Society of Arts.

Oller, J. and D. Obrecht. 1969. The psycholinguistic principle of information sequence: an experiment in second language learning. *IRAL 20*: 119-123.

Palmer, A. 1972. Testing communication. *IRAL 10*: 35-45.

———. 1978a. Compartmentalized and integrated control: an assessment of some evidence for two kinds of competence and implications for the classroom. Paper read at the 5th International Congress of Applied Linguistics, August, 1978, Montreal, Canada.

———. 1978b. Measures of achievement, communication, incorporation, and integration for two classes for formal EFL learners. Paper read at the 5th International Congress of Applied Linguistics, August, 1978, Montreal, Canada.

Paulston, C. 1970. Structural pattern drills: a classification. *Foreign Language Annals 4*: 187-193.

Rivers, W. 1969. From skill acquisition to language control. *TESOL Quarterly 3*: 3-12.

Upshur, J. 1969. Measurement of oral communication. In Schrand, H., ed., *Leistungmessung im Sprachunterricht*. Marburg/Lahn: Informationszentrum für Fremdsprachenforschung. 53-80.

Upshur, J. and A. Palmer. 1974. Measures of accuracy, communicativity, and social judgments for two classes of foreign language speakers. In *Selected papers from the Third International Congress of Applied Linguistics*, vol. 2. Heidelberg: Julis Gross Verlag. 201-221.

An Experiment in a Picture-Stimuli Procedure for Testing Oral Communication

Lyle F. Bachman
University of Illinois
at Urbana-Champaign

Abstract. As part of the evaluation of a five-year longitudinal research and development project in individualized language learning, several alternative methods for testing oral English production were tried out. The Bilingual Syntax Measure was selected for adaptation, because of the relative effectiveness of its visual component in eliciting responses. Adaptation of the test for native Thai-speaking upper elementary school children included modification of the content of the questions as well as the scoring procedure. A stratified random sample of 100 elementary grade 7 students were tested. Individual tests were tape-recorded, randomized, and prepared for rating. Raters included 5 native speakers of English and 1 native Thai-speaking English teacher. Both inter-rater correlations and internal consistency estimates of reliability were acceptable, while predictive validity correlations with measures of other language skills were highly significant. Content validity is claimed in that the test provides sufficient latitude for responses to go well beyond mere manipulation; questions require factual information about the pictures, inferences regarding causal relationships implied in the pictures, and inferences based on a common external frame of reference.

Background

The research reported in this paper was conducted as part of a five-year longitudinal research and development project aimed at developing and evaluating, in an experimental situation, the effectiveness of an individualized EFL program for upper elementary school in Thailand. (Aiken & Bachman, 1977) Both the experimental individualized program and the existing lock-step program with which it was compared included oral communication objectives and learning activities. But while the classroom evaluation procedures used in the two

programs were deemed adequate for assessing individual progress and achievement, they were not, because of the differences between the two programs, appropriate for use in assessing the comparative effectiveness of the two programs in teaching oral communication. It was therefore necessary to either adapt or develop a test of oral communication which could be standardized for use with both groups. That is, it was essential that both the content and the testing procedures be controlled so as to eliminate sources of bias to either program.

Try-Out

Initially, several distinct oral testing procedures were tried out with small groups of students comparable to those in the program. These procedures included a structured interview and adaptations of the Test of Spoken English (Baetens Beardsmore & Renkin, 1971; Baetens Beardsmore, 1974), PROTEST (Palmer, 1972), and the Bilingual Syntax Measure (Burt, Dulay, and Hernández-Ch., 1975). On the basis of this try-out, the Bilingual Syntax Measure (BSM) was selected for use in the program evaluation because of the relative effectiveness of its picture stimuli in eliciting responses and the degree of control of questions and scoring procedures its format permitted.

Adaptation

While some lexical changes had been made in the content of the BSM questions, it was apparent after the initial try-out that additional modifications would be necessary to eliminate a slight content bias that favored the individualized program. The try-out also revealed that several questions failed to elicit responses from students in either program. The adaptation of the test for pre-testing, therefore, included modifications in the content of the questions to better suit the content of the two curricula, and an increase in the number of questions to allow for item shrinkage after pre-testing.

The scoring procedure recommended for standard administrations of the BSM provides information on the grammaticality of subjects' responses. Since grammaticality was only one of the dimensions of oral communication to be evaluated, it was decided to supplement the BSM scoring procedure with ratings of fluency, pronunciation, grammaticality, and appropriateness.

Pre-Testing

Subjects for the pre-testing were 20 7th-grade students, 10 each from a school using the individualized curriculum and a school using the standard curriculum. These subjects were selected by stratified random sampling to reflect the widest possible range of language proficiency (2 high, 4 average, and 4 low subjects from each school, according to classroom teachers' assessment). Two

examiners, both native Thai-speaking members of the program staff, administered the test individually to individual students during regular class hours, in separate rooms, with subjects isolated upon completion of the test to minimize opportunities for test compromise. Subjects' responses were written down by the examiners, along with other information regarding performance on the test. All tests were also recorded on tape. The test tapes were edited, eliminating extraneous noise and long pauses between separate responses. From these edited tapes, 6 subjects were selected as representative of 6 levels of oral communication ability. One-minute segments of each of these 6 subjects' responses were selected, identified for proficiency level (1-6, with 6 being the highest), and prepared as a training tape. One-minute segments of all subjects' responses were selected and arranged at random on a rating tape. Raters included 5 native speakers of American English and 1 native Thai-speaking English teacher. These raters listened to the training tape at the beginning of the rating session, and were then asked to rate each of the 20 subjects on a scale from 1-6 for fluency, grammaticality, pronunciation, and appropriateness.

On the basis of the pre-test, 5 questions were eliminated as non-productive. No content changes were made in the remaining 25 questions. Two changes were made in the scoring procedure. Because of difficulties encountered by the raters in distinguishing appropriateness from grammaticality, and because of the bias introduced, in several cases, by the inclusion of the examiner's questions or repeated questions, it was decided to eliminate appropriateness as a factor to be rated on the final test, thus making possible deletion of all examiners' questions from the rating tapes. In order to provide a more finely differentiated scale of grammaticality than the 5 proficiency levels given by the BSM scoring procedure, the score of each subject was the total number of grammatically correct words uttered, following the criteria given in the BSM Manual.

Final testing

Subjects. One hundred 7th-grade students were selected at random, 50 each from classes using the individualized and standard curricula.[1] These students were from 8 through 12 years of age, with a mean age of 11, and had been studying English in a formal school setting for 5 hours per day, approximately 25 weeks per year, for 3 years. This sample included 53 female and 47 male students.

Procedures. The same examiners who administered the pre-test conducted the final testing, with each examiner testing 25 students from each curriculum group, in random order, following the same administrative procedures used for the pre-test. Training and rating tapes were edited as for the pre-testing, with the exception that the examiners' questions were edited out. The first two subjects

[1] Because of missing data on other variables in the study, this number was reduced to 95 for the final analysis.

on each rating tape were dummies — subjects from other schools, to be rated by the judges but not included in the analysis. As there were 5 rating tapes, with 20 subjects per tape, ratings were done in 5 separate sessions over a two-day period, with the training tapes being played at the beginning of each rating session. The same raters performed the ratings as in the pre-test.

Ratings were conducted as in the pre-test, with the 6 raters' ratings for each subject being combined to provide an average rating for each of the 3 factors, fluency, grammaticality, and pronunciation, as well as a rating total for overall oral communication. Subjects' tests were also scored to determine the total number of words grammatically correct, as in the pre-test.

Results

The reliability of the scoring procedures was estimated in two ways. Table 1 presents the inter-rater correlations among the 6 raters and the total rating score. The range of rater-total correlations was .608-.867, with an average correlation of .785 (using the z-technique for averaging). Internal consistency reliability estimates (KR_{21}) were also calculated. The two sets of scores were .737 and .984, for the rating total and words correct respectively.

TABLE 1
Inter-Rater Correlations

Rater	1	2	3	4	5	6	T
1	1.000						
2	.439	1.000					
3	.774	.515	1.000				
4	.713	.478	.704	1.000			
5	.517	.536	.552	.513	1.000		
6	.654	.725	.691	.680	.664	1.000	
T	.833	.720	.608	.823	.781	.867	1.000

(N.B. All correlations significant at $p << .001$, $df = 94$.)

Predictive validity was estimated by correlating the scores for oral communication with the scores for overall English, which consisted of a weighted average of scores on listening comprehension, reading comprehension, dictation, structure, and oral communication tests. Table 3 presents these correlations. The highest correlation is that between the Rating Total and Overall English scores. Furthermore, this correlation was significantly higher than that obtained between Words Correct and Overall English scores ($t = 9.89$, sig. at $p << .001$, $df = 92$).

TABLE 2
Correlations among Oral Communication Scores and Overall English Scores

	Overall English	Rating Total	Words Correct
Overall English	1.000		
Rating Total	.506	1.000	
Words Correct	.446	.426	1.000

(N.B. All correlations significant at $p << .001$, $df = 94$.)

Discussion

While the inter-rater reliabilities obtained with the oral communication test were not as high as those often reported for highly structured interviews with experienced examiners, they are within acceptable limits for a test of this length. The extremely high internal consistency estimate obtained for the Words Correct scores is an artifact of the extreme variation in these scores (Range: 0-143, S.D. = 36.68). This variation is almost certainly due as much to personality factors unrelated to oral communication as to variability in this skill itself. The KR_{21} estimate for the Rating Total scores (.737), however, is consistent with the average rater-total correlation (.785), since the KR_{21} formula normally underestimates reliability slightly.

The significantly higher correlation obtained between the Rating Total and Overall English scores suggests that the rating procedure provides more information than does the Words Correct scoring procedure. This is supported by a closer examination of these two procedures. As indicated above, the only criteria for correctness used in the Words Correct procedure relate to the grammaticality of the utterances. Indeed, to insure that other factors did not influence grammaticality judgments, this scoring was done from transcriptions of subjects' responses, rather than directly from the tapes themselves. The ratings, on the other hand, were made on the basis of segments of actual speech, so that judges were exposed to variations in pronunciation, fluency, and grammaticality, as well as a range of nonlinguistic signals indicating various states of nervousness, shyness, or interest.

The picture-stimulus question format of the test, while much more restricted than even a highly structured interview, nevertheless does provide sufficient latitude of responses to go well beyond mere manipulation. Although each question focuses on a specific picture-stimulus, questions range from yes-no and WH-questions requiring factual information about the pictures to questions requiring inferences. These require inferences within the context of the pictures (the fat man lives in the fat house), inferences regarding causal relationships implied in the pictures (the man isn't wearing shoes because he's mopping the deck of the ship), inferences based on a common frame of reference ("Why do

they want food?"), and sometimes inferences based upon imagination ("Why are the green fishes' eyes closed?"). Furthermore, the test comprises more than a series of isolated questions and answers. A number of questions, for example, depend on information provided in a previous response.

While the question format is flexible enough to allow creative communication in responses, the prior specification of the content and the number of the questions provides greater control over variability of subjects' responses. The form of the questions determines, to a large extent, the form and length of the likely responses, and thus helps control for wide differences in subjects' personalities and backgrounds. (The random selection of speech segments for rating also controls for this.) For this reason, less reliance for standardization needs to be placed on experienced examiners than is the case with oral interviews.

Problems of this testing procedure are primarily in the areas of development and scoring, and concern efficiency rather than reliability or validity. It is obvious that appropriate pictures and questions have to be developed for different groups. Here a major concern should be finding pictures that provide a rich enough context for the exchange of information, while avoiding the obvious pitfalls that introduce bias, cultural or otherwise, into the test content. Also important is the inclusion of questions that require creative input on the part of the respondent and that generate a context of discourse. Development of an appropriate form of the test thus involves trying the questions and pictures and analyzing the results, as outlined above. While this may be a negative feature in terms of efficiency, the fact that this procedure admits to this sort of analysis contributes to its reliability.

The most time-consuming aspect of the rating procedure is the editing and preparation of training and rating tapes. Indeed, without an experienced recording technician and adequate equipment, this is an insurmountable task. The rating sessions themselves, however, can be conducted quite efficiently. Furthermore, explicit instructions and representative training tapes virtually eliminate the need for experienced raters.

Conclusions

A picture-stimuli question format for testing oral communication provides results that are of acceptable reliability and which, it is argued, are valid in content. In addition, these results can be obtained through the use of standard administration and rating procedures, even without experienced examiners or raters. While the development of tests appropriate to specific groups is time-consuming, this involves procedures analogous to those regularly used in the development of more objective tests, and which permit item-banking.

REFERENCES

Aiken, P. and L. F. Bachman. 1977. *Individualizing EFL: curriculum research and development in Thailand*. Bangkok: Central Institute of English Language.

Baetens Beardsmore, H. 1974. Testing Oral Fluency. *IRAL 12*, 4: 317-326.

Baetens Beardsmore, H. and A. Renkin. 1971. A test of spoken English. *IRAL 9*, 1: 1-11.

Burt, M., H. Dulay and E. Hernández-Ch. 1975. *Bilingual syntax measure*. New York: Harcourt Brace Jovanovich.

Palmer, A. 1972. Testing communication. *IRAL 10*, 36-45.

APPENDIX
Bilingual Syntax Measure
ILLP Adaptation
Child Response Booklet

This booklet contains all the specific directions and questions for administering the BSM-E.

Child's name _____

Age: years _____ months _____ Boy _____ Girl _____

School _____ Grade _____

Date _____ Examiner _____

Notes and observations: (retest, special diagnosis, etc.)

_____

_____

_____

Show the child PICTURE 1 only. Then ask questions a. through e. in order

PRELIMINARY QUESTIONS (Do *not* record.)

a. **Do you see a fat man?. . . Show him to me.**
b. **And show me the thin man.**
c. **And the little birds up in the tree?**
d. Point to FLOWERS
And what are those?
e. Point to THE HAT
What is that?

TEST QUESTIONS (Record responses on lines provided.)

1. Point to LITTLE BIRDS
What are those? _____ 1. _____
2. Point to MOTHER BIRD and WORM
What's the mother bird going to do? _____ 2. _____
3. Point to LITTLE BIRDS
Why do they want food? _____ 3. _____
4. Point to FAT MAN
Why is he very fat? _____ 4. _____
5. Point to THIN MAN
Why is he very thin? _____ 5. _____

Show the child PICTURES 1 and 2 TOGETHER and say: **Let's look at another picture.**

PRELIMINARY QUESTIONS (Do *not* record.)

a. **Point to the fat house.** b. **And the thin house?**
c. **Where are the windows?** d. **And the doors?**

TEST QUESTIONS (Record responses on lines provided.)

6. Point to BOTH HOUSES using whole hand to point
What are these? _____ 6. _____
7. **What color is the fat house?** _____ 7. _____
8. Point to DOORS OF BOTH HOUSES AT ONCE
What are these? _____ 8. _____
9. Point to FAT MAN and FAT HOUSE
Why does he live here? _____ 9. _____

Now turn to the next picture and say: **Here's another picture!**
Show the child PICTURE 3 ONLY

PRELIMINARY QUESTION (Do *not* record.)

a. **Where are the fish?** b. **And the mop?** c. **And where are the man's shoes?**

TEST QUESTIONS (Record responses on lines provided.)

10. Point to MAN
What's he doing? _____ 10. _____
11. **Why is he doing that?** _____ 11. _____

12. **Why isn't he wearing his shoes?** _____ 12. _____
13. Point to the PAIL
What does the man have in the pail? _____ 13. _____
14. Point to EYES OF BOTH GREEN FISH
Why are the green fishes' eyes closed? _____ 14. _____
15. Point to EYES OF BOTH BROWN FISH
And why are their eyes open? _____ 15. _____
16. a. **What are the brown fish doing?** _____ 16a. _____
b. **What are the green fish doing?** _____ 16b. _____
17. a. **Is the man *all* wet?** _____ 17a. _____
b. **Why?** _____ 17b. _____
18. Point to MOP
Tell me, whose mop is that? _____ 18. _____
(If child just points, say "I didn't hear you.")

Now say to the child: **Here comes another picture!** And turn to the next picture.
Show the child PICTURE 4 ONLY.

TEST QUESTIONS (Record responses on lines provided.)

19. a. Point to GIRL
What's the girl doing? _____ 19a. _____
b. **Is she happy?** _____ 19b. _____
c. **Why?** _____ 19c. _____
20. Point to GIRL'S FLOWER
Whose flower is that? _____ 20. _____
(If child just points, say "I didn't hear you.")

Now say to the child: **Let's look at the last pictures**, and turn to the next pictures. Show the child PICTURES 5, 6, and 7 TOGETHER.

PRELIMINARY QUESTIONS (Do *not* record.)

a. Point to PICTURE 5
Where is the *king* in *this* picture?
b. Point to PICTURE 6
Where's the dog in *this* picture?
c. Point to PICTURE 7
And where's the king in *this* picture?:

TEST QUESTIONS (Record responses on lines provided.)

21. Point to DOG (PICTURE 5)
Why is the dog looking at the king? _____ 21. _____
22. Point to PLATE (PICTURE 7)
What happened to the king's food? _____ 22. _____
23. Point to PICTURE 6
Why did the dog take the king's food? _____ 23. _____
24. Point to PICTURE 7
Why is the king's plate empty? _____ 24. _____
25. Point to APPLE ON FLOOR (PICTURE 7)
Why did this apple fall down? _____ 25. _____

*A Multitrait-Multimethod Investigation into the Construct Validity of Six Tests of Speaking and Reading**

Lyle F. Bachman
University of Illinois, Urbana-Champaign
Adrian S. Palmer
University of Utah

Abstract. An empirical investigation into the construct validity of tests of speaking and reading English as a second language was performed using the multitrait-multimethod convergent-divergent design of Campbell and Fiske. Interview, translation, and self-rating tests of the two hypothesized traits, "speaking ability" and "reading ability," were administered to a population of 75 native speakers of Mandarin Chinese at the University of Illinois. The hypothesis of convergent validity was supported for all of the tests. The two hypotheses of discriminant validity were supported in enough instances to provide some evidence of this type of validity and, thus, evidence of the indepen-

*We want to acknowledge here that this study was a communal effort involving several institutions and many individuals. The CIA Language School provided the facilities and personnel to train us in administering the FSI oral interview test. The FSI School of Language Studies invited us to observe tests and advised us on test administration and development procedures. And our own institutions, the University of Illinois and the University of Utah, provided us with funding and released time to conduct the study.

The participants in the 1979 Boston colloquium spent two days in formal meetings and many hours outside deciding upon a preliminary research design, which was refined during the months that followed in endless phone conversations and exchanges of letters. Randall Jones and Harold Madsen provided us with the questionnaire used to obtain demographic information on the subjects. Pardee Lowe and Ray Clifford spent four days teaching us as much as we could absorb about the intricacies of the oral interview test — as fine a training program as one could wish. George Trosper, one of Palmer's graduate students, helped immeasurably in the development of the reading tests. Several graduate research assistants at the University of Illinois were also instrumental in the project's completion. Jennifer Lin and Lilia Wang, MATESL students there, contacted subjects, provided translations of all tests and correspondence, assisted in test development, and administered and scored the reading tests. Don Anderson, also a MATESL student, organized the testing schedule and administered the self-ratings and the recorded oral translation exam. Steve Dunbar, a Ph.D. student in educational evaluation, was instrumental in coding, processing, and analyzing the data.

dence of the speaking and reading traits. An analysis of variance was also performed which supported the hypothesis that speaking and reading abilities are independently measurable. In addition, it provided evidence that the method of testing has a significant influence on the test scores. The Campbell-Fiske design for collecting data is endorsed, but newer ways of formulating and testing the hypotheses used in evaluating the data are advocated.

Introduction

One goal of the 1979 Boston colloquium was to stimulate empirical research into the construct validity of tests of communicative competence. The plan adopted by the participants was to proceed in two phases. In Phase 1, evidence as to the construct validity of tests of global areas of language use was to be sought. If evidence of such validity was found, Phase 2, an investigation into the construct validity of the *components* of communicative competence, would be undertaken.

We were among the researchers present at the colloquium, and undertook to carry out Phase 1 of the investigation. This paper describes the study as actually performed[1] and presents an interpretation of the results based essentially on the Campbell-Fiske criteria described in the Introduction to this volume and in the paper by Clifford.

The steps of our procedure were as follows: 1) defining traits and selecting methods, 2) operationalizing the definitions of trait-method units in the form of tests, 3) stating hypotheses, 4) administering and scoring the tests, and 5) evaluating the hypotheses in light of the results. It will be seen that these are the steps of the general procedure given in the section on "the construct validity of oral proficiency tests" in the Introduction, slightly modified to make them applicable to the multitrait-multimethod (MTMM) design. The description of the study in this paper follows this sequence of steps except that, for readability, the statement of the hypotheses (step 3 above) is delayed in order to present the hypotheses concurrently with their evaluation (step 5).

[1] During the early planning stages of this project, this Phase 1 study came to be referred to among the colloquium participants as "The Quick-and-Dirty Pilot Study," and was viewed as a mere preliminary exploration preceding the main study. The latter was to be of truly monumental proportions — subjects totalling in the tens of thousands, testing sites throughout the world, a million-dollar budget, etc. (Yes, those were the numbers actually bandied about.) Carrying out the Phase 1 study quickly snapped us back into The Real World. It was hardly quick. Some 160 man-hours went into instrumentation, 40 hours into contacting subjects, four days (and a trip to the CIA Language School in Washington) into training us to administer the FSI oral interview test, 280 man-hours (five examiners working seven 8-hour nonstop days each) into administering the tests, 260 man-hours into rating, scoring, and coding the data, and 200 man-hours into programming and analysis time. The final bill came to approximately $30,000, including computer time. So much for the quickness. As for the dirt, we feel that the amount of time spent planning the study in collaboration with many generous expert researchers contributed to our obtaining remarkably "clean" data — no missing data, highly reliable scores, etc.

Defining Traits and Selecting Methods

Our decision as to which traits to investigate in this Phase 1 study was influenced primarily by our desire to select two maximally distinct aspects of language competence, thus maximizing the probability of our finding more than one trait — if more than one did, in fact, exist. Therefore, we decided to investigate tests of the hypothesized traits "global competence in speaking" and "global competence in reading," traits differing both in channel (aural versus visual) and in direction (production versus reception).

We chose for our trait definitions the FSI global descriptions of "absolute language proficiency" in speaking and reading. These descriptions characterize proficiency at 11 levels (0, 0+, 1, 1+, . . . , 5). These are given in the Appendix to this paper.

These particular trait descriptions were selected because the FSI scales, particularly in their application to the FSI oral interview, 1) are described in detail in the literature, 2) are widely used, and 3) have become the subject of considerable interest and controversy.

The methods selected were an interview method, a translation method, and a self-rating method. The choice of the interview and translation methods was again influenced by the high level of general interest in the FSI proficiency tests, which use an interview method to measure speaking proficiency and a translation method to measure reading proficiency. The self-rating method was chosen because it was easily adapted to the measurement of both the speaking and the reading traits. Several other methods had been proposed and discussed over a period of several months. One, a multiple-choice paper-and-pencil method, seemed practical and was of considerable general interest, but we were unable to devise a way of testing the speaking trait via this method which was even face valid. We felt that for this Phase 1 study we should use only methods which at least appeared fairly well-suited to the measurement of the hypothesized traits. The traits, methods, and resultant tests finally chosen are shown in Figure 1 and discussed further below.

Operationalizing the Definitions of the Trait-Method Units (Tests)

The interview test of speaking

The CIA version of the FSI oral interview was selected. It consists of a face-to-face interaction definitively described by Lowe (1976a, 1976b) which is designed to elicit a sample of the testee's speech ratable using the FSI trait definitions. Unlike the FSI's own version of the method, it does not purport to measure listening comprehension directly.

The researchers, Bachman and Palmer, were put through a four-day intensive training program at the CIA Language School by Pardee Lowe and Ray

FIGURE 1
Tests used

Traits \ Methods	Interview (1)	Translation (2)	Self Rating (3)
Speaking (A)	FSI Oral Interview (CIA version which does not attempt to get at aural comprehension directly).	The subject is asked to translate replies to questions or directives written in his native language into spoken English and to record his translation. These replies vary in complexity according to the FSI absolute language proficiency descriptions.	The subject rates his own speaking ability on a scale similar to that used by FSI examiners.
Reading (B)	An interview, conducted in the subject's native language. The subject reads passages at various levels selected according to FSI procedures. The examiner asks the subject both general and detailed questions about the meaning of the passages. The subject responds in his native language. The answers to the questions do not require direct translation from English to the subject's native language.	The FSI reading test, administered *not* as an interview, but as follows: the subject is given a set of graded passages in English to translate line by line into his native language.	The subject rates his own reading ability on a scale similar to that used by FSI examiners.

Clifford. The testing procedures were discussed, interviews were observed, and practice interviews were administered and criticized.

The interview test of reading

An interview format was developed for testing reading comprehension. In this test, the subject was given a short passage in English to be read silently. When ready, he was asked a number of questions about the passage in his native

language (Chinese, in this study), which he answered, also in his native language. The questions were of the following five types, none of which required the subject to translate directly from the English passage into Chinese.

1) Questions to be answered by pointing to information in the passage
2) Yes-no questions
3) Questions asking for a summary of part or all of the passage
4) Questions requiring comprehension of particular words or phrases
5) Questions requiring comprehension of the organization of the passage

The passages were selected according to the criteria for oral translation passages set out in the FSI's *Testing Kit* (FSI, 1979: 41-44). These specify the types of sources to be used at the five FSI levels. For level 1, we produced a list consisting of individual signs of one to three words and/or numeral groups, such as those one would encounter on a street or in a building, and a short passage of the type found in beginners' language textbooks. The level-2 reading was adapted from a junior high school science magazine. The newspaper item for level 3 was taken from the news columns of the *New York Times*. One level-4 passage was chosen from the instructions to an income tax form and the other was a handwritten copy of a humorous letter by Jean Kerr. For level 5, three passages were necessary: a very formal essay by Cardinal Newman, a comic piece by Phyllis Diller, and the handwritten text used at level 4 (with different questions).

The translation test of speaking

The translation test of speaking was constructed by adapting the unpublished Recorded Oral Production Examination (ROPE) developed by Ray Clifford and Pardee Lowe. The ROPE consists of a set of recorded questions or directives at FSI levels 1-4. Question types at each level follow the guidelines set out in Lowe (1976). In the original ROPE, the subject listens to the question and is given time to respond. The tape is then rated as per the FSI guidelines.

For this study, the ROPE test was converted into a record oral translation examination (ROTE) by supplying the testee with a written Chinese version of an answer to the recorded question/directive. This answer was designed to elicit, when retranslated, English grammatical structures and lexical items consistent with the descriptions of competence at the FSI level for which the eliciting question/directive was prepared. Thus, the ROTE test as used consisted of the following steps:

1) The subject listened to a tape recording on which he heard a question or directive in Chinese. At the same time, he read the question/directive, also in Chinese.
2) He was given a period of time to read an appropriate response in Chinese.

3) Upon signal, he then translated the response into English, recording it on tape.
4) This procedure was repeated for all of the questions at each of the four levels.

The translation test of reading

The procedure used by the FSI for testing reading was slightly adapted. Though called an interview, the FSI reading test is essentially a translation test. In it, the subject sits down with two examiners. Based usually on their knowledge of the subject's proficiency in speaking, they give him a short reading passage in the language to be tested and assign all or part of it to be translated orally into the subject's native language. The examiners occasionally supply obscure vocabulary items or request the subject to repeat; but except for the face-to-face postures of subject and examiners, interaction is slight or nonexistent. The FSI examiners rate the translation and, depending on its adequacy, assign another passage at the same level or at a higher or lower level. The process is repeated until a final rating is assigned.

In the test as used in this pilot study, the interaction between examiner and subject during the translation was minimized even further. The test administrator conducted a brief interview in Chinese to determine what kind of material the subject read in English, and what his occupation and educational background were. Based on this, the administrator assigned the first passage. She then supervised the tape recording of the subject's translation and determined the level of the additional passage or passages assigned.

Since the FSI itself does not test proficiency in reading English as a second language by means of their test, we selected English passages according to their criteria (FSI, 1979: 41-44). The passages used were generally different sections from the same sources used in the interview test of reading; the few exceptions were of very similar type and difficulty (e.g., a piece by I. A. Richards on literary criticism corresponded to the passage by Cardinal Newman in the interview test).

The self-rating tests

The self-rating tests of reading and speaking were written questionnaires in Chinese. Each contained two basically different types of questions. One type probed the subject's perception of his functional control of spoken and written English. In these questions, he was asked what he could *do* with the language — what language use situations he could cope with. The situations were based on the functional portions of the FSI global descriptions of absolute language proficiency (see Appendix) and on the FSI's own self-appraisal questionnaire reflecting those descriptions (FSI, 1979: 18-22). The second type of question

probed the subject's perception of his general control of linguistic forms (range and accuracy). These levels of control were also drawn from the FSI descriptions of the five levels of competence. The questions were grouped according to FSI level.

Pretesting

All tests were informally pretested on small groups of native Mandarin speakers who were excluded from the study itself. Test procedures and items were modified as required.

Administering and Scoring the Tests

Sample

In order to facilitate administration of tests involving translation, it was decided early in the study to sample subjects from a homogeneous native-language background. Given the objectives of this Phase 1 study, it was felt that any possible loss in generality of findings was outweighed by practical considerations of data gathering. Therefore, a group of native Mandarin Chinese-speaking students at the University of Illinois, Urbana-Champaign, was identified. Subjects were contacted at random, using a list of Chinese students obtained from the International Student Office, University of Illinois. In order to increase the variability of the sample, student spouses were also asked to participate. Eighty-five subjects were scheduled for testing and sent background information questionnaires. Of these 85, 4 did not appear for testing, 2 were eliminated because their control of Mandarin was not sufficient for them to complete the translation tests, and 4 were eliminated because they missed one of the tests. Subjects were paid $5.80 each for their participation.

The sample thus comprised 75 native Mandarin Chinese speakers from Taiwan who were living in Illinois. 61 were university students (57 graduate, 4 undergraduate) majoring in 39 different fields, 13 were spouses of students, and 1 was enrolled in an intensive English institute. There were 39 females and 36 males, ranging in age from 19 to 35 years, with a median age of 26 years. 25 had been living in the U.S. for less than one year, while 50 had been living in the U.S. for one year or more. All had studied English for at least one year on Taiwan, and 61 had studied English for more than one year here. 30 indicated that they knew languages other than Chinese and English (French-5; German-10; Japanese-13; Spanish-2; and Malay-1). Of these, only one indicated a better knowledge of speaking and reading that language (Malay) than English.

Administrative procedures

Each subject took all six tests in sequence, in a single two-hour period. Subjects were scheduled at half-hour intervals, at their convenience. Tests were administered individually by project staff.

The two researchers, Bachman and Palmer, administered the oral interviews. The two reading tests (interview and translation) were administered by two native Mandarin-speaking students, each of whom was seeking a master's degree in teaching English as a second language (MATESL). The self-ratings and the Recorded Oral Translation Examination (ROTE) were administered by a native English-speaking MATESL student.

Rating and scoring procedures

Each of the two interviewers administering the oral interview assigned an independent rating (0-5 FSI scale) to each subject immediately upon completion of the interview, after which a joint "conference" rating was assigned for use in the analysis. For the reading interview and reading translation tests, each interviewer assigned an independent FSI rating to each subject and then rated tapes of the other's sessions, thus providing two sets of ratings for each measure. From these an average rating for each subject was computed for use in the analysis. The tape recordings of the ROTE were also rated independently by two raters (Bachman and Palmer, in this case) and average ratings computed for use in the analysis. Scores for the two self-appraisals were the total number of questions answered "yes" by each subject on each measure.

Evaluating the Hypotheses in Light of the Results

Analyses

Distributions, correlations and reliabilities were computed using SPSS Version 8 on the CYBER system at Illinois. Multiple analysis of variance was computed using the method described by Stanley (1961).

Reliability estimates

Because of the effect of attenuation on correlations, the estimation of reliability is crucial to inferences to be made from the multitrait-multimethod (MTMM) matrix. This is stated explicitly by Campbell and Fiske as the first criterion in the MTMM inference structure. In order to allow disattenuation of the correlations obtained among the six trait-method units, reliabilities were estimated using variance components of the scores. For the ratings (oral interview,

reading interview, reading translation, and ROTE), the intraclass correlation was used, and for the self-rating, Guttman's lambda 6, a lower bounds estimate (Guttman, 1945) was used. Both of these estimates are compatible with the assumptions made for disattenuation regarding sources of error, and for both only the assumption of independent variance (raters or items) need be made. (The assumption of equal variance — homogeneous item or rater difficulty — is not necessary.) In addition to these estimates, which are of prime interest for analyzing the MTMM matrix, both inter- and intra-rater reliabilities were estimated to determine the stability of the ratings across raters and across time. The obtained reliabilities are given in Table 1.

TABLE 1
Reliability Estimates for Trait-Method Units

	Oral Interview	Reading Interview	ROTE	Reading Translation	Speaking Self-rating	Reading Self-rating
Inter-rater (N = 75)	.887	.974	.849	.943	NA	NA
Intra-rater (N = 30)	—	.984	—	.997	NA	NA
Intra-class (N = 75)	.878*	.974*	.860*	.944*	NA	NA
Alpha (N = 75)	NA	NA	NA	NA	.908	.851
Lambda 6 (N = 75)	NA	NA	NA	NA	.959*	.894*

NA = Not appropriate
— = Not computed
* = Used in correcting for attenuation, and reported in Table 3

Multitrait-multimethod (MTMM) correlations

The MTMM correlation matrix, corrected for attenuation, is given in Table 2. Correlations marked with "R" are reliability estimates. Those marked with "C" are indicators of convergent validity, and those marked with "M" and "H" are used to assess different aspects of discriminant validity. All correlations above and to the right of the solid line are duplicates of values found elsewhere in the table; these are included only to facilitate finding the values to be used in the testing of hypotheses 3 and 4.

TABLE 2
MTMM Correlation Matrix, Corrected for Attenuation, with Reliabilities on the Diagonal
(All correlations significant at $p<.01$, df = 74.)

		Speaking (A)			Reading (B)		
		Int (1)	Tran (2)	Self (3)	Int (1)	Tran (2)	Self (3)
A	1	.88R	.90C	.57C	.53M	.63H	.51H
	2	.90C	.86R	.57C	.74H	.77M	.51H
	3	.57C	.57C	.96R	.62H	.51H	.74M
B	1	.53M	.74H	.62H	.97R	.69C	.73C
	2	.63H	.77M	.51H	.69C	.94R	.60C
	3	.51H	.51H	.74M	.73C	.60C	.89R

Several specific hypotheses regarding the reliability, convergence, and discrimination of the trait-method units were tested in the MTMM framework, and several inferences can be made. These are presented below, with evaluations in light of the results.

Hypothesis 1: σ_e^2 = (approx.) 0

Random error variance is near zero.

This implies high reliabilities (near 1.00) for all trait-method units.

Since the lowest obtained reliability is .86, while the highest is .97, this hypothesis is supported.

Hypothesis 2: (Convergence) $C > 0$

Monotrait-heteromethod correlations (C) should be significantly higher than zero, and "sufficiently large to encourage further examination of validity." (Campbell and Fiske, 1959: 33)

High correlations between different methods for measuring the same trait are seen as evidence of convergent validity; low monotrait-heteromethod correlations indicate lack of convergence and preclude further examination of discriminant validity.

The correlations in the lower right-hand triangle (reading triangle) of Table 3 converge quite well, with values of .73, .69, and .60. In the speaking triangle in the upper left-hand corner, the interview and translation methods converge very well (.90), while the convergence of the self-rating with the other measures is

lower (.57 in both cases) but still statistically significant. These results support the hypothesis of convergence.

Hypothesis 3: (Discrimination) C > H
Convergent validity coefficients (C) should be higher than the correlations (H) between tests having neither trait nor method in common.

This hypothesis is tested by comparing each of the validity coefficients (labeled C in the broken-line triangles) with the correlations between tests having neither trait nor method in common with each other (those labeled H) but which each share *either* trait *or* method (but not both) with the validity coefficient in question. Four comparisons will be examined for each validity coefficient C.

For example, if we wish to evaluate the discriminant validity of tests of speaking, we compare the convergent validity coefficients for the speaking tests (in the upper left-hand triangle) with those correlations marked H which fall within the same column or row as the coefficient in question. (H values in the same column will share trait, but not method, with the C value; H values in the same row will share method, but not trait, with the C value.) Thus, we would compare the validity coefficient .90 with the H values .63 and .51 in the column below it, and with the H values .74 and .51 in the row to its right.

Discriminant validity of tests of speaking. There are twelve relevant comparisons involving the validity coefficients for the speaking tests: .90 with .63, .51, .74, .51; .57 with .63, .51, .62, .51; and .57 with .74, .51, .62, .51. The validity coefficients here are higher than the H values in 7 out of 12, or 58%, of the cases.

Discriminant validity of tests of reading. There are also twelve relevant comparisons involving the validity coefficients for the reading tests: .69 with .63, .51, .74, .62; .73 with .51, .51, .74, .62; and .60 with .51, .51, .63, .51. The validity coefficients here are higher than the H values in 9 out of 12, or 75%, of the cases.

Summary. Hypothesis 3 is supported in 16 out of 24, or 67%, of the cases, providing some evidence of discriminant validity.

Hypothesis 4: (Discrimination) C > M
Convergent validity coefficients (C) should be higher than the correlations obtained between different traits measured by the same method (M).

Intuitively, high heterotrait-monomethod correlations would indicate dominance of method, and hence invalidate the test. Low heterotrait-monomethod correlations are interpreted as additional evidence of discriminant validity.

In evaluating the evidence for hypothesis 4, the monomethod correlations (M) are compared with validity coefficients (C) in the same column or row. If

method effect is low (which is necessary if we are to find evidence of discriminant validity), these monomethod correlations should be lower than the relevant convergent validity coefficients.

Discriminant validity of tests employing the interview method. There are four relevant comparisons for the validity coefficients of tests using the interview method: .53 with .90, .57, .69, .73. The monomethod correlation is lower than all the convergent validity coefficients, supporting hypothesis 4 in 100% of the cases where the interview is the method of testing used.

Discriminant validity of tests employing the translation method. There are four relevant comparisons for the validity coefficients of tests using the translation method: .77 with .90, .57, .69, .60. The monomethod correlation is lower than one of the four convergent validity coefficients, supporting hypothesis 4 in only 25% of the cases where translation is the method of testing used.

Discriminant validity of tests employing the self-rating method. There are four relevant comparisons for validity coefficients of tests using the self-rating method: .74 with .57, .57, .73, .60. The monomethod correlation is lower than none of the four convergent validity coefficients, providing no support for hypothesis 4 where self-rating is the method of testing used.

Corroboration. This pattern of greater method effect for the interview and self-rating methods than for the interview method can also be observed if we compare the three monotrait-monomethod correlations with each other. (This is not, strictly speaking, part of the direct test of hypothesis 4). When the interview method is used to measure both speaking and reading, the correlation between test scores is .53. In contrast, when the translation method is used, the correlation is .77, and when the self-rating method is used, it is .74. This suggests that the interview and self-rating methods exert a greater influence on test scores than does the interview.

Summary. Hypothesis 4 is supported only for the interview method.

Analysis of variance (ANOVA) of the MTMM matrix

The limitations of directly comparing pairs or sets of correlations in the MTMM matrix have been discussed in a number of studies (Jackson, 1969; Boruch et al., 1970; Alwin, 1974; Kalleberg and Kluegel, 1975) and center primarily around the problem of explicitly distinguishing and quantifying the variance due to different main effects and interactions. For example, comparisons between convergent validities and the corresponding heterotrait-monomethod values in Table 2 above suggest that there is a subject-method interaction, but provide no means for quantifying or determining the significance of this interaction.

In one approach to this problem, Stanley (1961) has shown that MTMM data can be analysed by analysis of variance (ANOVA), treating the MTMM design as a three-way factorial with subject, trait, and method as factors. Several

studies using this approach, with subject as a random effect and trait and method as fixed effects, have demonstrated the advantages of ANOVA in interpreting MTMM data (Boruch et al., 1970; Kavanaugh et al., 1971; Mellon and Crano, 1977). These advantages are generally that data can be summarized and interpreted more efficiently, particularly with large matrices, and that validity information is more explicit and quantifiable. Specifically, with regard to the explicitness of validity information, two advantages accrue: (1) the magnitude of the differences among main effects and interactions with respect to change can be appraised, and (2) the relative magnitudes of the component variances conditional on the model can be estimated (Boruch et al., 1970; 841).

Within the ANOVA framework, the hypotheses pertaining to convergent and discriminant validity are as follows:

Hypothesis 1: $MS_{subject} > 0$

A significant MS (mean square) for this main effect (as indicated by the corresponding F value) reflects general agreement among the six trait-method units in measuring the subjects and is indicative of convergent validity.

Hypothesis 2: $MS_{subject \times trait} > 0$

A significant MS for this interaction indicates significant measured differences among subjects on different traits and reflects the differential meaning of the two traits, i.e., discriminant validity.

Hypothesis 3: $MS_{subject \times method} > 0$

A significant MS for this interaction reflects the bias of some methods towards certain subjects and indicates the amount of method effect on subjects' scores. A non-significant MS for this interaction indicates the absence of method effect and is further evidence of discriminant validity.

The results of the three-way ANOVA are presented in Table 3.

TABLE 3
Analysis of Variance of Correlations

Source	df	SS	MS	F
Subject	74	314.955	4.256	18.393**
Subject × trait	74	37.620	0.508	2.184**
Subject × method	148	63.180	0.427	1.845++
Error	148	34.245	0.231	—

**sig. at p < .01, df = 74, 148
++sig. at p < .01, df = 148, 148

The significant main effect for subjects shown in the table indicates strong convergence of the trait-method units (tests). That is, there is general agreement

among the six tests in ranking the subjects across traits and methods. The significant subject-trait interaction reflects the amount of variance due to unique traits, and suggests a degree of independence between the two traits examined. The significant subject-method interaction reflects the differential effect of method across subjects.

Thus, the analysis of variance indicates that the most important effect on the scores is attributable to differences among the subjects. Regardless of how they were tested or what they were tested on, subjects tended to be ordered rather similarly. Next in importance was the unique effect of the traits measured — the reading and speaking traits. In other words, while there is considerable similarity in the ordering of the subjects across traits and methods, significant trait differences also exist. This suggests the presence of a general trait in addition to the speaking and reading traits. Finally, of almost equal importance with the effect of the trait is the unique effect of the method of measurement used: interview, translation, or self-rating.

Conclusion

This study has clearly shown that performance on language tests is influenced by at least two independent factors: the effect of test method and the effect of the trait(s) being measured. It has shown that the effect of method on test scores is stronger for translation and self-rating than for the interview method. And it has provided some support for the hypothesis that there are independently measurable speaking and reading traits — enough support to warrant continuing with Phase 2 of the investigation. We believe that these three conclusions exhaust those to be drawn from the data, as considered from the perspective of the Campbell-Fiske hypotheses and the analysis of variance.

With respect to methodology, we note that the Campbell-Fiske procedure can be divided into two parts: a general design for collecting data, and a set of hypotheses for evaluating the data collected. We have found no way of improving on the Campbell-Fiske design for collecting data. However, we maintain that any study examining constructs should work within a model that allows the researcher to quantify the effect of test method on test scores. This is only imperfectly possible when the Campbell-Fiske hypotheses are used. Our reading in the psychometric literature during the time we were analyzing our data has led us to alternate ways of formulating and testing hypotheses which we believe are more powerful and enlightening than those discussed in this volume and used in this paper. (See our forthcoming papers in which we recommend confirmatory-factor-analytic procedures.) We suggest that future studies examining construct validity (including Phase 2 of this investigation) should, while incorporating the Campbell-Fiske data collection scheme, frame and test hypotheses according to these procedures.

REFERENCES

Alwin, Duane F. 1974. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costar, ed., *Sociological methodology 1973-1974*. San Francisco: Jossey-Bass.

Althauser, R. P., T. A. Heberlein, and R. A. Scott. 1971. A causal assessment of validity: the augmented multitrait-multimethod matrix. In H. M. Blalock, ed., *Causal models in the social sciences*. Chicago: Aldine-Atherton.

Bachman, L. F. and A. S. Palmer. Forthcoming. The construct validity of the FSI Oral Interview. *Language Learning*, June 1981.

_____. Forthcoming. Construct validation of foreign language tests: methodological considerations. In Dougias Stevenson, ed., *Advances in language testing: series 4*. Washington, D.C.: Center for Applied Linguistics.

Boruch, R. F., J. D. Larkin, L. Wolins, and A. C. MacKinney. 1970. Alternative methods of analysis: multitrait-multimethod data. *Educational and Psychological Measurement 30*: 833-853.

Campbell, D. T. and D. W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin 56*, 2: 81-105.

Foreign Service Institute (FSI). n.d. *Testing kit: School of Language Studies*. Washington, D.C.: Department of State.

_____. 1979. *Testing kit: French and Spanish*. Washington, D.C.: Department of State.

Guttman, L. 1945. A basis for analyzing test-retest reliability. *Psychometrika 10*, 4: 255-282.

Jackson, D. N. 1969. Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin 72*, 1: 30-49.

Kalleberg, A. L. and J. R. Kluegel. 1975. Analysis of the multitrait-multimethod matrix: some limitations and an alternative. *Journal of Applied Psychology 60*, 1: 1-9.

Kavanaugh, M. J., A. C. MacKinney, and L. Wolins. 1971. Issues in managerial performance: multitrait-multimethod analysis of ratings. *Psychological Bulletin 75*, 1: 34-49.

Lowe, Pardee, Jr. 1976a. *Handbook on question types and their use in LLC oral proficiency tests*. Preliminary version. Washington, D.C.: CIA Language Learning Center.

_____. 1976b. *The oral language proficiency test*. Washington, D.C.: Interagency Language Roundtable.

Mellon, P. M. and W. D. Crano. 1977. An extension and application of the multitrait-multimethod matrix technique. *Journal of Educational Psychology 69*, 6: 716-723.

Stanley, J. C. 1961. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika 26*, 2: 205-219.

APPENDIX
FSI Global Definitions of Absolute Language Proficiency in Speaking and Reading*

The rating scales described below have been developed by the Foreign Service Institute to provide a meaningful method of characterizing the language skills of foreign service personnel of the Department of State and of other Government agencies. Unlike academic grades, which measure achievement in mastering the content of a prescribed course, the S-rating for speaking proficiency and the R-rating for reading proficiency are based on the absolute criterion of the command of an educated native speaker of the language.

The definition of each proficiency level has been worded so as to be applicable to every language; obviously the amount of time and training required to reach a certain level will vary widely from language to language, as will the specific linguistic features. Nevertheless, a person with S-3's in both French and Chinese, for example, would have approximately equal linguistic competence in the two languages.

The scales are intended to apply principally to government personnel engaged in international affairs, especially of a diplomatic, political, economic and cultural nature. For this reason heavy stress is laid at the upper levels on accuracy of structure and precision of vocabulary sufficient to be both acceptable and effective in dealings with the educated citizen of the foreign country.

As currently used, all the ratings except the S-5 and R-5 may be modified by a plus (+), indicating that proficiency substantially exceeds the minimum requirements for the level involved but falls short of those for the next higher level.

Elementary Proficiency

S-1 *Able to satisfy routine travel needs and minimum courtesy requirements.* Can ask and answer questions on very familiar topics; within the scope of very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak the language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at the S-1 level should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.

R-1 *Can read simplest connected written material, authentic or especially prepared for testing.* In a form equivalent to usual printing or typescript, can read either representations of familiar verbal exchanges or simple language containing only the highest frequency grammatical patterns and vocabulary items. Texts may include personal and place names, street signs, shop designations and office designations.

Limited Working Proficiency

S-2 *Able to satisfy routine social demands and limited work requirements.* Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on nontechnical subjects (i.e. topics which require no specialized knowledge) and has a speaking vocabulary sufficient to respond simply with some circumlocutions; accent, though often quite faulty, is intelligible;

* *From Foreign Service Institute (1979: 13-15), except for the R-5 definition. That definition, apparently by printing error, does not appear in FSI (1979) and has been supplied from the previous edition, FSI (n.d.: 15).

can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.

R-2 *Can read simple authentic written material in a form equivalent to usual printing or typescript on subjects within a familiar context.* Can read uncomplicated but authentic prose on familiar subjects such as news items describing frequently occurring events, simple biographic information, social notices, formatted business letters and simple technical material written for the general reader. The prose is predominantly in familiar sentence patterns. Test candidates may need occasional prompting on low frequency items.

Professional Proficiency

S-3 *Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversation on practical, social, and professional topics.* Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.

R-3 *Able to read standard newspaper items addressed to the general reader, routine correspondence, reports and technical material in own special field.* Can grasp the essentials of articles of the above types without using a dictionary; for accurate understanding moderately frequent use of a dictionary is required. Has occasional difficulty with unusually complex structures and low-frequency idioms.

Distinguished Proficiency

S-4 *Able to use the language fluently and accurately on all levels normally pertinent to professional needs.* Can understand and participate in any conversation within the range of own personal and professional experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

R-4 *Able to read all styles and forms of the language pertinent to professional needs.* With occasional use of a dictionary can read moderately difficult prose readily in any area directed to the general reader, and all materials in own special field including official and professional documents and correspondence; can read reasonably legible handwriting without difficulty.

Native or Bilingual Proficiency

S-5 *Speaking proficiency equivalent to that of an educated native speaker.* Has complete fluency in the language such that speech on all levels is fully accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

R-5 *Reading proficiency equivalent to that of an educated native.* Can read extremely difficult and abstract prose, as well as highly colloquial writings and the classic literary forms of the language. With varying degrees of difficulty can read all normal kinds of handwritten documents.