DOCUMENT RESUME

ED 186 45[illegible] TM 800 142

AUTHOR Bachman, Lyle F.; Palmer, Adrian S.
TITLE Convergent and Discriminant Validation of Oral Language Proficiency Tests.
PUB DATE [Sep 79]
NOTE 11p.: Paper presented at the International Conference on Language Proficiency and Dominance Testing (3rd, Carbondale, IL, September 26-28, 1979).
AVAILABLE FROM University of Illinois, 3070 Foreign Languages Building, Urbana, IL 61801 ($0.50)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Communicative Competence (Languages); *English (Second Language); *Evaluation Methods; Higher Education; Interviews; Language Proficiency; *Language Tests; Mandarin Chinese; Oral Reading; Reading Comprehension; Research Design; Self Evaluation (Individuals); *Speech Skills; *Test Validity; Translation
IDENTIFIERS Multitrait Multimethod Techniques

ABSTRACT

In a study designed to validate oral language proficiency tests, it is planned to administer a series of tests to 100 native Mandarin Chinese-speaking subjects (foreign students and their spouses). The tests will measure communicative competence in speaking (ability to speak, exhibiting control of linguistic, sociolinguistic, and pragmatic rules; and fluency) and communicative competence in reading (ability to react to these rules as manifested in written language, and to react fluently). Three different testing methods will be used, resulting in a multitrait-multimethod design: interviews, translation, and self-rating. The results will verify hypotheses of competence, and the components of the construct, oral proficiency. (Author/GDC)

Convergent and Discriminant Validation of Oral Language Proficiency Tests

Lyle F. Bachman
University of Illinois, Urbana-Champaign

Adrian S. Palmer
University of Utah

Recently, considerable research has been devoted to testing oral language proficiency, and a number of different oral testing procedures have emerged (Clark, 1975, 1978, 1979; Jones, 1975, 1979; Palmer and Groot, 1979). Central to much of this research is the acceptance of "face validity" as a criterion for evaluating oral proficiency tests, and the reliance on concurrent validation procedures for relating "indirect" to "direct" testing methods.

Both of these approaches to validity have been shown to be of dubious utility (Cronbach and Meehl, 1955; Stevenson, 1979). The problems inherent in criterion-referenced validation include not only the difficulty of establishing an adequate criterion measure, but also the potentially serpentine process of successive approximation. While circular validation procedures are generally precluded by conscientious test developers, the problem of valid criterion measures remains. The proposed solution to this problem in the area of language proficiency testing has been the appeal to the "face validity" of so-called "direct" measures (Clark, 1975, 1979; Jones, 1975). The notion of "face validity" in the case of language proficiency is intuitively appealing. Obviously the most direct sample of speaking proficiency, for example, is for someone to speak. That is, the most direct sample of a given behavior is the behavior itself. This becomes less obvious, however, when we consider another mental trait, intelligence. No one, I believe, would claim that digit-symbol substitution, which is one test in the Wechsler Adult Intelligence Scale, can be equated with intelligence. That is, we do not posit the identity of a trait with its behavioral manifestation. The problem of "face validity," then, as it has been advocated in language proficiency testing, is that it confuses the outward manifestation of a trait with the trait itself. Once this is recognized, the distinction between "direct" and "indirect" measures becomes irrelevant, since all tests sample manifestations of traits, and not the traits themselves. With this recognition that all tests are indirect measures of traits, the notion of "face validity" becomes "the mere appearance of validity, [and] is not an acceptable basis for interpretive inferences from test scores." (APA, 1974, p. 26). Therefore, the claim that a given test is a valid measure of a given trait cannot be accepted on the basis of an appeal to "face validity," but must be supported by a much more rigorous inference structure, that of construct validation.

In construct validation, or the process of investigating what psychological constructs are measured by a given test, a test is validated not against another test, but against a theory. To investigate construct validity, one develops a construct (a theory) which becomes a provisional explanation of test results until the theory is falsified by the results of testing hypotheses derived from it. Thus the test becomes, in essence, the operational definition of the construct.

The model of construct validation to be followed in this study is the multitrait-multimethod matrix (Campbell and Fiske, 1959). This design recognizes that any test score is a function of both the trait it intends to measure and of the method by which it is measured. In order to distinguish trait variance from method variance in test scores, the design requires that at least two distinct traits be measured by at least two separate methods. High correlations between scores on different measures of the same trait would demonstrate convergent validity. High correlations between similar methods of measuring different traits, however, would invalidate the test, and so we interpret low correlations between such measures as evidence of discriminant validity. In the multitrait-multimethod matrix, method variance can thus be delineated from trait variance, so that both convergent and discriminant validity can be examined.

The basic research design of this study is a 2x3 multitrait-multimethod matrix, with the following traits and methods given below. Trait definitions are derived from the general framework developed by Canale and Swain (1979).

Traits

A. Communicative competence in speaking consists of:

1. The ability to produce spoken language exhibiting control of the linguistic rules employed by speakers of a given dialect or set of dialects. Control consists of breadth (the range of structures attempted) and accuracy (the degree to which structures are produced correctly). The areas of linguistic control are syntax, phonology, and lexicon.

2. The ability to produce spoken language exhibiting control of the sociolinguistic rules employed by speakers of a given dialect or set of dialects. Sociolinguistic rules consist of the conventions for producing speech in a register appropriate to specific speech situations. Control consists of breadth (the range of speech situations in which the speaker is sensitive to different prevailing standards) and accuracy (the degree to which the language produced conforms to prevailing standards).

3. The ability to produce spoken language which exhibits control of the pragmatic rules employed by speakers of a given dialect or set of dialects for communicating the types of messages required by these speakers. Pragmatic rules are conventions for relating the form of an utterance to its intended meaning. Control consists of breadth (the range and complexity of messages communicated) and accuracy (the degree to which the language produced communicates correctly the details of the content).

4. The ability to produce spoken language fluently. Fluency consists of quickness of response to perceived needs to speak and the rate of speech (the degree to which the tempo of the speech conforms in overall speed and consistency of speed to norms for speakers of a given dialect or set of dialects).

B. Communicative competence in reading consists of:

1. The ability to react to the linguistic rules manifested in written language. Ability consists of breadth (the range of structures reacted to) and accuracy (the degree to which reactions are correct). Areas of linguistic control are graphology, syntax and lexicon.

2. The ability to react to the sociolinguistic rules employed in a given written dialect or set of dialects. Sociolinguistic rules consist of conventions appropriate to particular aims and modes of written discourse. Control consists of breadth (the range of aims and modes in which the speaker is sensitive to prevailing standards) and accuracy (the degree to which reactions conform to prevailing standards).

3. The ability to react to the pragmatic rules employed in a given written dialect or set of dialects to communicate types of messages appropriate to that dialect or set of dialects. Pragmatic rules are conventions for relating the form of a text to its intended message. The ability to react consists of breadth (the range of messages) and accuracy (the degree to which the reactions conform to prevailing standards).

4. The ability to react to written language fluently. Fluency consists of the rate of response to written material (the degree to which responses conform to norms for readers of a given dialect or set of dialects).

Methods

A. The interview method consists of a face-to-face language use situation requiring subjects to interact with one or more interlocutors, exchanging information, but requiring no direct translation from the subjects' native language to the target language or vice-versa.

B. The translation method consists of a language use situation requiring the subjects to translate directly from their native language to the target language and/or vice-versa. The situation is not face-to-face and there is no interaction with an interlocutor.

C. The self-evaluation method consists of subjects' self-ratings in their native language of their ability in the specified traits in the target language. There is no use of the target language, no direct translation, and no interaction. This design can be schematized as in Figure 1 below.

	Method 1 (Interview)	Method 2 (Translation)	Method 3 (Self-evaluation)
Trait A (Oral proficiency)	A_1	A_2	A_3
Trait B (Reading proficiency)	B_1	B_2	B_3

Figure 1
Trait-method units in the multitrait-multimethod design of the study

Trait-method unit A_1 will consist of a standard FSI-type oral interview. This highly structured interview consists of several distinct parts and includes well-defined procedures for checking the subjects' levels of proficiency and for probing to determine the upper bounds of these levels. (Wilds, 1975; Lowe, 1976). Research indicates that this testing procedure has high reliability (Adams, 1978; Clifford, 1978; Mullen, 1978), predictive validity (Clark, 1975; Jones, 1978), and high concurrent validity (Shohamy, 1979; Hendricks et al., 1979). Interviews will be conducted with two interlocutors. Simultaneous ratings, both individual and conference, will be given. In addition, interviews will be tape-recorded, arranged in random order, and rated at a later time.

Trait-method unit A_2 will consist of a series of short dialogues in the subjects' native language which they will listen to and then provide a direct oral translation into English. These dialogs will vary in ways consistent with the types of control specified in the definition of speaking. The translations will be recorded on tape, arranged in random order, and rated, using a scale consistent with the FSI rating scale.

Trait-method unit A_3 will consist of the subjects' self-rating of their oral proficiency in English. Both Lickert and semantic-differential scales will be used.

Trait-method unit B_1 will consist of a set of graded reading passages in English, which the subjects will read. An interlocutor will then ask questions about the passage in the subjects' native language, which they will answer in their native language. The discussion will focus on the subjects' comprehension of the reading passage, and no direct translation will be required. The interview will be recorded and rated at a later time.

Trait-method unit B_2 will consist of a set of graded reading passages in English which the subjects will read silently and then translate directly, line by line, into their native language. These translations will be tape recorded and rated at a later time.

Trait-method unit B_3 will consist of the subjects' self-rating of their reading proficiency in English. Both Lickert and semantic-differential scales will be used.

Subjects for this study will be limited to a homogeneous mother-tongue group of non-native speakers of English. The intended sample of 100 subjects will be drawn by stratified random sampling of Mandarin Chinese-speaking students and spouses of students at the University of Illinois.

The basic analytic procedure for multitrait-multimethod data is to examine the interrelationships among the various trait-method units in the matrix. One method for doing this is to compute a matrix of correlations, as illustrated in Figure 2[1].

---

[1]While correlation is the basic analysis proposed by Campbell and Fiske, three other analytic procedures have been proposed: analysis of variance (Mellon and Crano, 1977), confirmatory factor analysis (Jöreskog, 1969; Kalleberg and Kluegel, 1975) and consistency criteria (Althauser,

	Traits	Method 1		Method 2		Method 3	
		A_1	B_1	A_2	B_2	A_3	B_3
Method 1	A_1	r					
	B_1	d_m	r				
Method 2	A_2	c	d_h	r			
	B_2	d_h	c	d_m	r		
Method 3	A_3	c	d_h	c	d_h	r	
	B_3	d_h	c	d_h	c	d_m	r

Figure 2

Idealized 2x3 multitrait-multimethod matrix

r = reliabilities (monotrait-monomethod correlations)

c = convergent validities (monotrait-heteromethod correlation

d_m = discriminant correlations (heterotrait-monomethod)

d_h = discriminant correlations (heterotrait-heteromethod)

In this matrix, we can identify a diagonal row of correlations between the same method and the same trait. These monotrait-monomethod correlations (r) comprise the reliabilities of the trait-method units involved. Two other diagonals parallel to the reliability diagonal consist of correlations between different methods of measuring the same trait. These monotrait-heteromethod correlations (c) comprise convergent validity coefficients. The two other sets of correlations are 1) those between different methods of measuring different traits (heterotrait-heteromethod--d_h) and 2) those between the same method for measuring different traits (heterotrait-monomethod--d_m).

The logic of inferences to be made from multitrait-method data requires two assumptions: First, we assume that random error variance approaches zero. This assumption necessitates high reliabilities for all tests. If this requirement is not met, subsequent inferences from the data are highly questionable. The second assumption pertains to

---

Heberlein and Scott, 1971). These procedures, in addition to the examination of correlations described by Campbell and Fiske, will be considered, particularly in the follow-up studies suggested below.

non-random error variance, and is two-fold: method and trait factors are uncorrelated, and method variance is constant across traits. (Alwin, 1974). These latter assumptions must be met if inferences regarding discriminant validity are to be valid.

The hypotheses and inferences from multitrait-multimethod data are as follows:

1. $c > 0$

   Monotrait-heteromethod correlations (c) should be significantly higher than zero, and "sufficiently large to encourage further examination of validity." (Campbell and Fiske, p. 33) High correlations between different methods for measuring the same trait are seen as evidence of convergent validity. Low monotrait-heteromethod correlations (c) indicate lack of convergence and preclude further examination of discriminant validity.

2. $c > d_h$

   Convergent validity coefficients (c) should be higher than the correlation obtained between different methods for measuring different traits (d_h). Low heterotrait-heteromethod correlations (d_h) are interpreted as evidence for discriminant validity.

3. $c > d_m$

   Convergent validity coefficients (c) should also be higher than the correlations obtained between different traits measured by the same method (d_m). Intuitively, high heterotrait-monomethod correlations (d_m) would indicate dominance of method, and hence invalidate the test. Low heterotrait-monomethod correlations are interpreted as additional evidence of discriminant validity.

4. Similar patterns of trait interrelationships in all heterotrait groups. For example, if the rank-order of correlations in one grouping is $c > d_h > d_m$, we would expect to find the same order in other such groupings.

Within this framework, the hypotheses of this project pertain to the following general questions regarding language proficiency:

1. Is there evidence that the trait "communicative competence in speaking" is distinct from the trait "communicative competence in reading"?

2. If these traits are distinct, what are the components of communicative competence in speaking?

3. If these traits are not distinct, what factor(s) is/are common to both?

The hypothesis of distinct traits (speaking and reading) will be supported if the data show evidence of both convergent and discriminant validity. Specifically, if we find 1) high correlations among the three methods on the same traits and 2) lower correlations among different methods of measuring different traits and among the same methods for measuring different traits, we will have evidence to support the hypothesis of distinct skills. In this case, the analysis of additional ratings of pronunciation, grammar, fluency and vocabulary will be included in the matrix as traits, and analyzed to determine their distinctness. If only convergent but not discriminant validity is evidenced, then the hypothesis of distinct skills will not be supported. In this case, the analysis of additional ratings may lead to hypotheses regarding factors common to both speaking and reading. Such analyses will provide more precise definitions of the traits examined. These definitions in turn will form the basis for hypotheses of subsequent research into the components of communicative competence.

In this paper, we have argued that criteria currently followed for evaluating the validity of language proficiency tests are inadequate. We have presented a specific model and set of procedures for investigating both convergent and discriminant validity. In the study presently being conducted, a widely accepted and used procedure for testing oral language proficiency, the FSI oral interview, will be examined for construct validity. A definition of oral proficiency based on a model of communicative competence is proposed as a framework for stating hypotheses. The results of this study will bear upon the unitary factor hypothesis of language proficiency. Further research will investigate the components of communicative competence, both in separate skill areas and in general.

Bibliography

Adams, M. L. 1978. Measuring foreign language speaking proficiency: a study of agreement among raters, in J.L.D. Clark, ed. Direct testing of speaking proficiency: theory and application. Princeton: Educational Testing Service.

Alwin, Duane F. 1974. Analyzing the multitrait-multimethod matrix. In H.L. Coster, ed. Sociological methodology 1973 - 1974. San Francisco: Jossey-Bass.

Althauser, Robert B. 1974. Inferring validity from the multitrait-multimatrix: another assessment. In H.L. Coster, 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-multimehtod matrix, in H.M. Blalock, ed. Causal models in the social sciences. Chicago: Aldine - Atherton.

American Psychological Association, 1974. Standards for educational and psychological tests and manuals. Washington: American Psychological Association.

Brière, B. and F. Hinofotis, 1979. Concepts in language testing: some recent studies. Washington: Teachers of English to Speakers of Other Languages.

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

Canale, M. amd M. Swain, (forthcoming). A theoretical framework for communicative competence, in Palmer, A.S. amd P.J.M. Groot, eds. The validation of oral proficiency tests. Washington, D.C.: Teachers of English to Speakers of Other Languages.

Clark, J. 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.

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

Clark, J. L. D. 1979. Direct vs. semi-direct tests of speaking ability, in E. Briere and F.B. Hinofotis, eds. Concepts in language testing: some recent studies. Washington, D.C.: Teachers of English to Speakers of Other Languages.

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

Cronbach, L. J. 1971. Test validation. In R.L. Thorndike, ed. Educational measurement, 2nd Ed. Washington: American Council on Education.

Cronbach, L. J. and P. E. Meehl. 1955. Construct validity in psychological tests. Psychological Bulletin 52, 4.

Hendricks, D. et al. (forthcoming). Three pragmatic tests of oral proficiency and the FSI oral interview: an evaluation. In A.S. Palmer and P.J.M. Groot, eds., The validation of oral proficiency tests: an introduction, Washington: Teachers of English to Speakers of Other Languages.

Jackson, D. N. 1969. Multimethod factor analysis in the evaluation of convergent and discriminant validity. Psychological Bulletin 72, 1.

Jones, R. L. 1975. Testing language proficiency in the United States government in R.L. Jones and B. Spolsky, eds. Testing language proficiency. Arlington, VA: Center for Applied Linguistics.

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

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

Jöreskog, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis, Psychometrika 34.

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.

Lowe, P., L., Jr. 1976. The oral language proficiency test. Washington: 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.

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. Direct testing of speaking proficiency: theory and application. Princeton: Educational Testing Service.

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

Palmer, A. S. and P. J. M. Groot, eds. (forthcoming). The validation of oral proficiency tests: an introduction. Washington: Teachers of English to Speakers of Other Languages.

Shohamy, E. (forthcoming). Inter-rater and intra-rater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew. In A.S. Palmer and P.J.M. Groot, eds. The validation of oral proficiency tests: an introduction. Washington: Teachers of English to Speakers of Other Languages.

Stevenson, D. K. (forthcoming). Beyond faith and face validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests. In A.S. Palmer and P.J.M. Groot, eds. The validation or oral proficiency tests: an introduction. Washington: Teachers of English to Speakers of Other Languages.

Wilds, C. P. 1975. The oral interview test. In R.L. Jones and B. Spolsky, eds. Testing language proficiency. Arlington, VA: Center for Applied Linguistics.