ED 023 077

AL 001 516

By-Spolsky, Bernard

Language Testing. The Problem of Validation.

Teachers of English to Speakers of other Languages.

Pub Date Jun 68

Note-7p., Paper presented at the TESOL Convention, San Antonio, Texas, March 1968.

Available from-TESOL, Institute o f Languages and Linguistics, Georgetown University, Washington, D.C. 20007 ($1.50).

Journal Cit-TESOL Quarterly: v2 n2 Jun 1968

EDRS Price MF-$0.25 HC-$0.45

Descriptors-*English (Second Language), *Language Tests, *Test Validity

Foreign language tests fall into two classes, according to their purposes. The first class, tests used for the control of instruction, may be achievement or diagnostic tests. The second class of tests, used in the control of a person's career, may be concerned with what the subject can do, or what he should be able to do in the future. The temporal distinction is less important than the major functional one: exactly the same test can serve as a diagnostic test before some material i s taught, and as an achievement test after. Similarly, proficiency tests are generally used as predictors of future performance. The author suggests using a functional definition of levels which would test ability to operate easily and effectively in specified sociolinguistic situations (rather than attempting to characterize levels of knowing a language in terms of grammatical and lexical mastery). As in all testing, the central problem of foreign language testing is validity. With tests of the first class, this problem is not serious, for the textbook or syllabus writer has already specified what should be tested. With tests of the second class, it remains a serious difficulty, for no way has yet been found to characterize knowledge of a language with sufficient precision to guarantee the validity of the items included or the type of tests used. (AMM)

Language Testing—The Problem of Validation[*]

Bernard Spolsky

Foreign-language tests fall naturally into two classes according to the purposes for which they are used. In the first class are tests used for the control of instruction. They are the concern of the classroom teacher who wishes to find out how effective his teaching and the students' learning have been or to discover what needs to be taught. The second class are tests used in the control of a person's career. Used by administrators or counsellors, they are intended to help make decisions about someone's qualifications for a given task or about the type of training he should follow.

Each of these classes may be further divided according to the temporal relation of the test to its goal. Tests of the first class concerned with testing what has been taught are achievement tests; those concerned with what is about to be taught are diagnostic tests. Tests of the second class concerned with what the subject can do now are achievement tests; those concerned with what he should be able to do in the future are predictive tests. But this temporal distinction is less important than the major functional one; exactly the same test can serve as a diagnostic test before some material is taught and as an achievement test after it. Similarly, proficiency tests are generally used as predictors of future performance.

This functional classification agrees with one that can be made on operational grounds: tests of the first class are relatively simple to prepare and straightforward to interpret, while tests of the second class involve serious theoretical and practical difficulties in preparation, interpretation, and especially, in validation. Why this is so becomes clear if we consider the steps to be followed in preparing a test.

Take the preparation of a test of the first type, an achievement or diagnostic test, to be used by a classroom teacher either before she starts a unit or chapter in the textbook or after she has finished. The starting point is the syllabus, with its list of items to be learned in the unit. The purpose of the test will be to decide how many of the items on the list have been mastered by the students. For our example, let us assume that we have an elementary class in English as a second language; we wish to test their knowledge of vocabulary, and our syllabus is defined by Lesson Eleven of Book One of *English for Today*.[1] Notice that the first, and in many ways, most important task of test writing has been done: the syllabus (or textbook in this case) gives us a list of the sixteen new words in the lesson. There is no point in our going

[*] This paper was presented at the TESOL Convention, March 1968.

Mr. Spolsky, presently Assistant Professor of Linguistics at Indiana University, is the associate editor with Paul Garvin of *Computation in Linguistics: A Case Book* (Bloomington: Indiana University Press, 1966). In September, 1968, Mr. Spolsky moves to the University of New Mexico where he will be Associate Professor of Linguistics and Elementary Education.

[1] *English for Today*, ed. William R. Slager (New York: McGraw-Hill, 1962).

beyond the list unless of course we wish to test words previously taught. From it, we select the words to be tested. If we have time, we can put every word in the test, but there is no need, for using some appropriate techniques, we can choose a representative sample. Next, we have to decide on the testing technique we are going to use. It is here that we are called on to define more precisely what it means to "know vocabulary"; we need to translate the general term into a more precise one. Here are some possible operational definitions, each describing a possible technique:

(a) When presented with a word on the list, the student taking the test should say, "I know it" or "I don't know it."

(b) When presented with a word on the list, he must select which one of a group of definitions is appropriate:

glass— something you drink out of
something you paint with
something you draw with

(c) When presented with a picture of the object named by a word on the list he indicates its name:

glass
cup
bottle

(d) When presented with a picture, the student must write down what it is.

There are of course many other techniques,[2] but these may be considered a representative sample. Of course, they each raise minor problems. The first, (a), might not be considered a normal sort of test, but it is likely to be a most useful technique with teachers and students who are cooperating closely in the learning process. The second and third raise a special problem: when multiple choice items are given, the student should know the meaning of the incorrect answers as well as the correct one; otherwise, the choice is unduly limited. For example, in (c) above, *bottle* would be a bad distractor, for the word is not introduced in Lesson Eleven (or in fact in Book One). Similarly, (b) has a bad definition; the word *drink* comes in Lesson Sixteen; in Lesson Eleven, all you do with glasses is wash them. A more serious problem in choosing a test technique is deciding whether it is a valid representation of the skill we want to test. Is there a serious difference between being able to recognize a definition and being able to give a definition? The former technique is easy to mark, the latter takes much longer. But it is quite easy to try out all the different techniques, and decide for ourselves whether they correlate so well that we only need to use one in the future.[3] Once we have decided

[2] See, for details, Robert Lado, *Language Testing* (New York: McGraw-Hill, 1961), or Rebecca Valette, *Modern Language Testing: A Handbook* (New York: Harcourt, Brace and World, 1967).

[3] *The Interpretive Information for the Test of English as a Foreign Language* (Educational Testing Service, 1967, revised January 1968) for example describes an interesting comparison of the scores on the "writing" section (a set of multiple-choice questions) with the scores of the same students on a set of essays graded by a team of examiners. The correlation is .74, which is close enough to suggest that the saving in time is worthwhile, unless of course we are planning to interpret the scores as though they had 100 percent validity. And on this see Paul Holtzman's paper in *NAFSA Studies and Papers*, English Language Series, Number 13: ATESL Selected Conference Papers (1967).

on the items and technique, the rest of the task of preparation is simple. And interpretation is straightforward, too. As long as the test is a representative sample of items, its result will say, "This student scored sixty percent on the test; he knows sixty percent of the words on the list." If the test is a diagnostic test, we will know what words need be taught; if an achievement test, we will know how effective our teaching has been. What has made test preparation and interpretation so simple has been that we have been able to ask a question to which there is a quantifiable answer. We have not asked whether or not the student knows English vocabulary, but rather how many of the words on this list he knows. Our results are clear, for they say he knows a given percentage of the words in Lesson Eleven of the textbook.

Basic to this relative simplicity was the existence of a list of items to be tested. Clearly, such lists are not available for all tests used in control of instruction. But it is equally clear that effective teaching depends on the availability of clear specifications. Normally, we have a syllabus or textbook or both, with lists of vocabulary, grammatical structures, etc. With such a syllabus or textbook, test making is straightforward. A control of instruction test is concerned with the question, "Have the items listed in the syllabus or textbook been learned to some criterion level?" It is not concerned with what should be learned. It would be wrong to include in a test of this class items that are not included in the syllabus.

When we say then that an achievement test is not a good one, we are referring to its inability to test a defined body of material; we are not saying anything about what should constitute that material. That is the task of the syllabus or textbook writer. Now, there are clearly cases when the distinction between test writer and textbook writer are blurred. One such case is when the test writer is trying to evaluate achievement in something that has not in fact been specified. He then has to do the textbook writer's job of specification before he can prepare an achievement test. This happens when one has a set of materials that can be listed as items and patterns, but one wishes to test the ability of the students to know more than the items or patterns they have been taught. For example, one may wish to find out about a student's ability to speak naturally on a topic other than those he has been trained for in memorized dialogues, or to use patterns with words other than those included in the pattern drills. But in such cases, we are really moving out of the realm of achievement tests, and into the area of proficiency, the second class of tests. These cases in fact set the limit; the first class in its purest form consists of tests defined not only functionally but also operationally—functionally, in that they are used in the control of instruction, and operationally, in that they are tests prepared on the basis of specifications of behavior or items that have been prepared, independently of the test, as part of the development of materials, textbooks, and syllabus.

The second class of foreign-language tests is defined functionally as tests used primarily in the control of a subject's career. They serve to make

dgments possible on such questions
:

1. How well will the subject do learning foreign languages in gener-, or one foreign language in particur? Should he be advised (permitted, ncouraged) to study a language? hould his employer (or the governent, or the armed forces) invest time nd money in his studying a lanuage?

2. How well does the subject perorm in the given foreign language? f he needs to use the language in overnment or other service, will he e successful? If he is a graduate stuent in a given field, will he be able o read books in the foreign language?

Tests aimed to handle the first set of uestions are predictive tests; their ask is to make some sort of judgment ossible on the question of the stuent's language-learning aptitude, and will need to make available information on any factors that will be relevant to language learning. This type of test sets many basic problems about the nature of second language acquisition, but will be left out of consideration in this paper.[4] Here, we shall be concerned with tests intended to answer questions of the second sort, proficiency tests.

Fundamental to the preparation of valid tests of language proficiency is a theoretical question of what it means to know a language. There are two ways in which this question can be answered. One is to follow what John Carroll[5] has referred to as the integrative approach, and to accept that there is such a factor as overall proficiency. The second is to follow what Carroll called the discrete-point approach: this involves an attempt to break up knowing a language into a number of separate skills, and further into a number of distinct items making up each skill. We are using the overall approach when we give a subjective evaluation of the proficiency of a foreign speaker of our language. In such cases, we usually do not refer to specific items that he has or hasn't mastered but to his ability to function in a defined situation. We do not say, "He is unable to distinguish between the phonemes /i/ and iy/," but rather something like "He doesn't know enough English to write an essay, but he seems to be able to follow lectures and to read his textbooks without much trouble." The key assumption of the discrete-point approach is that it is possible to translate sentences of the second type into a list of sentences in the first, and the key requirement for discrete-point testing is that we could quantify "He knows the words on this list."

Detailed instructions on how to prepare tests like this are given in the books by Lado and Valette referred to earlier. Drawing in particular on the powers of techniques developed by taxonomic linguistics to describe in detail the surface structure of languages, Lado shows how it is possible to construct tests that permit very fine discrimination of the strengths and weaknesses of foreign-language

[4] For a discussion of this problem, see Paul Pimsleur, "Testing Foreign Language Learning," *Trends in Language Teaching*, ed. Albert Valdman (New York: McGraw-Hill, 1966).

[5] John B. Carroll, "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students," *Testing* (Center for Applied Linguistics, 1961).

learners. Basic to Lado's approach is a theory calling for systematic description of the surface structure of the language being learned, combined with comparison with the language of the learner; it leads to a notion that tests as well as teaching material should be based on contrastive analysis, and prepared accordingly. Using these techniques, it is possible to develop tests that may be scored objectively (although some studies have raised some questions about the type of question used)[6] and the results of which lead to such precise interpretation tion as "the subject confuses medial /l/ and /r/." Tests of this nature are obviously of very great value in the control of instruction, whether as diagnostic or achievement tests.

But we must ask whether such an approach, assuming that all we have to do is to list all the items, permits us to characterize overall proficiency. If so, overall proficiency could be considered the sum of the specific items that have been listed and of the specific skills in which they are testable. To know a language is then to have developed a criterion level of mastery of the skills and habits listed. There are rather serious theoretical objections to this position. First, a discrete-point approach assumes that knowledge of a language is finite in the sense that it will be possible to make an exhaustive list of all the items of the language. Without this, we cannot show that any sample we have chosen is representative and thus valid. We must then argue for selection on the basis of functional necessity. This involves defining the functional load of the ability to distinguish between a pair of phonemes or of the ability to recognize the appropriateness of a given verb form. To do this, we would have to collect a list of minimal pair utterances in which the distinction is vital, but there turn out to be very few real minimal-pair situations, that is, situations where a single linguistic difference in a given situation will lead to complete misunderstanding. I have been told for instance the true story of a foreign lady speaking to her Italian maid: she asked to have the meat (*carne*) brought to the table, but had it given to the dog (*cane*) instead; rather strong punishment for speaking an r-less dialect. The rarity of such situations is a result (and theoretical cause) of the redundancy of natural languages.[7] Thanks to redundancy, we can communicate satisfactorily without knowing any given item in a language. This is most obvious in the area of vocabulary, where it is quite clear how many of the words in the dictionary are unknown to the average native speaker; it is true also in the area of phonology, otherwise speakers of different dialects of the same language would never be able to understand each other. It is probably not true in the case of many syntactic rules, but many of these are likely to turn out to be universal, and so irrelevant to foreign-language testing. More important, though, is the fact that syntactic rules are untestable unless fleshed out with vocabulary and phonology or spelling.

[6] See for instance, Eugène Brière, "Testing the Control of Parts of Speech in FL Compositions," *Language Learning*, XIV, 1 & 2 (1964).

[7] For a brief account of this, see John B. Carroll, *Language and Thought* (Englewood Cliffs: Prentice-Hall, Inc., 1964).

All of this suggests the impossibility of characterizing levels of knowing a language in linguistic terms, that is, as mastery of a criterion percentage of items in a grammar and lexicon. A more promising approach might be to work for a functional definition of levels: we should aim not to test how much of a language someone knows, but test his ability to operate in a specified sociolinguistic situation with specified ease or effect. The preparation of proficiency tests like this would not start from a list of language items, but from a statement of language function; after all, it would not be expected to lead to statements like "He knows sixty percent of English," but "He knows enough English to shop in a supermarket."

Functional statements of language proficiency may take various forms. One of the most thorough examples of a fairly complete scale is that prepared by the Foreign Service Institute for the classification of officers of the U.S. State Department. These Absolute Language Proficiency Ratings, as they are called, involve a division into language skills (reading, writing, speaking, and comprehending) and a numerical rating for each. The numerical ratings are generally described by a brief title, and range from "elementary," through "working" and "professional" to "native or bilingual." For each level of each skill there is a short description, again emphasizing skill. For example, to receive the rating S-3, one must be "able to speak the language with sufficient structural accuracy and vocabulary to satisfy representation requirements and handle professional discussions within a special field." There is then a longer description, suggesting the type of language-learning experience that is associated with the level.[8]

Starting with functional statements of this sort (and there should be little problem in preparing such description for each of the situations in whicı. proficiency tests are used), the language tester's problem is to find a reliable, valid, and economical method of rating a subject's proficiency in these terms. The first question is one of strategy. The discrete-point approach implies that it is pυssible to give a linguistic description of each level, to list the words and grammar needed to achieve this, but this is not possible either in theory or practice. The practical approach followed in the past has been to decide in some ad hoc way (the opinion of teachers, for instance) on the sort of items to be tested and the sort of test to use, but even though such a test can be made extremely reliable, it proves impossible to show its validity with sufficient precision to justify interpretations or improvements.[9] A more helpful strategy is to prepare proficiency tests in two stages. For the first stage one must forget considerations of expense and time. Expensive tests, using panels of trained judges, and having the subject function in situations of the sort

[8] The Absolute Language Proficiency Ratings are described in a number of mimeographed circulars. More accessible is the sample quoted by John Carroll in his article in *Foreign Language Annals* I, 2, (December 1967), and the description by Frank Rice in the *Linguistic Reporter* (May 1959).

[9] This problem has been discussed, among other places at a seminar held at the 1967 Conference of the National Association for Foreign Student Affairs, the proceedings of which have been published in *ATESL Selected Conference Papers.*

described in the rating scales, should first be developed as yardsticks. The second stage then involves taking cheaper procedures, of whatever kind, and correlating them with the more expensive measures. The degree of correlation will show the value of the ad hoc tests and make clear the degree of doubt that must be kept in their interpretation.

The exact nature of these more practical tests is not important; one would presume that they would be similar to many of the tests presently being used,[10] but they would permit of greater confidence in use, greater possibility of improvement (for we could then be in a position to speak about improving the validity of an objective test), and greater refinement in interpretation. It is probable that we would be able to develop simpler tests (e.g., the overall proficiency test using redundancy I have been working on[11]), and so ultimately justify the expense of the validation procedures.

The central problem of foreign-language testing, as of all testing, is validity. With tests of the first class, used by classroom teachers in the control of instruction, this problem is not serious, for the textbook or syllabus writer has already specified what should be tested. With tests of the second class, it remains a serious difficulty, for we have not yet found a way to characterize knowledge of a language with sufficient precision to guarantee the validity of the items we include or the types of tests we use.

[10] John Carroll, for instance, has investigated the correlation between the FSI Absolute Languag@ Proficiency Ratings and the MLA Foreign Language Proficiency Tests for Teachers and Advanced Students.

[11] Bernard Spolsky, Bengt Sigurd, Masahito Sato, Edward Walker, and Catherine Arterburn, "Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency," *International Review of Applied Linguistics*, (in press)..