ED 020 515 AL 001 292

CRITERION-REFERENCED TESTING OF LANGUAGE SKILLS.

BY- CARTIER, FRANCIS A.

PUB DATE MAR 68

EDRS PRICE MF-$0.25 HC-$0.40 8P.

DESCRIPTORS- *ENGLISH (SECOND LANGUAGE), *LANGUAGE SKILLS, *TESTING, *INSTRUCTIONAL TECHNOLOGY, SECOND LANGUAGE LEARNING, TESTS, COURSE OBJECTIVES, CURRICULUM DEVELOPMENT, CRITERION TESTS,

THE AUTHOR DISCUSSES THE CONCEPTS OF INSTRUCTIONAL TECHNOLOGY AS THEY APPLY TO THE PROBLEM OF TEACHING ENGLISH AS A FOREIGN LANGUAGE. INSTRUCTIONAL TECHNOLOGY, AN OUTGROWTH OF PROGRAMMED INSTRUCTION, HAS GROWN TO HAVE A FAR GREATER BREADTH OF APPLICATION AND MAY REPRESENT AN EVEN MORE FUNDAMENTAL CHANGE OF INSTRUCTIONAL PHILOSOPHY THAN PROGRAMMING. ITS MOST IMPORTANT RAMIFICATIONS HAVE LITTLE TO DO WITH INSTRUCTIONAL MEDIA OR METHODS, BUT MORE WITH DETERMINATION OF COURSE OBJECTIVES AND WITH EVALUATION OF WHETHER THE STUDENTS HAVE ACHIEVED THOSE OBJECTIVES. AN OUTSTANDING FEATURE OF THE PROCEDURE SUGGESTED HERE IS THAT THE INSTRUCTIONAL TECHNOLOGIST STARTS BUILDING HIS CURRICULUM BY PREPARING THE FINAL EXAMINATION, AND THEN BUILDS A COURSE THAT TEACHES THE STUDENT TO PASS THE EXAMINATION. THE TEST DOES NOT MERELY SAMPLE PARTS OF THE COURSE, BUT COVERS EVERYTHING THE STUDENT MUST LEARN TO DO, AND EVERY STUDENT IS EXPECT TO GET EVERY ITEM RIGHT. THE AUTHOR CONTRASTS THE "CRITERION TEST" WITH THE TRADITIONAL KIND OF NORM-REFERENCED TESTS AND DESCRIBES ITS APPLICATION AT THE DEFENSE LANGUAGE INSTITUTE'S ENGLISH LANGUAGE SCHOOL. THIS ARTICLE APPEARS IN "TESOL QUARTERLY," VOLUME 2, NUMBER 1, MARCH 1968, PUBLISHED BY TEACHERS OF ENGLISH TO SPEAKERS OF OTHER LANGUAGES, AT THE INSTITUTE OF LANGUAGES AND LINGUISTICS, GEORGETOWN UNIVERSITY, WASHINGTON, D.C. 20007. (AMM)

TESOL QUARTERLY

Volume 2 March, 1968 Number 1

Table of Contents

TESOL QUARTERLY

A Journal for Teachers of English to Speakers of Other Languages

Membership in TESOL ($6.00) includes a subscription to the journal.
TESOL QUARTERLY is published in March, June, September, and December.
Business correspondence should be addressed to James E. Alatis, Institute of Languages and Linguistics, Georgetown University, Washington, D.C.

Criterion-Referenced Testing of Language Skills

Francis A. Cartier

Five or six years ago, the term *instructional technology* was introduced into the professional jargon of the Air Training Command and, within a year or two, could be seen in Army and Navy training publications as well. The term was an outgrowth of programmed instruction, but has grown to have a far greater breadth of application and perhaps represents an even more fundamental change of instructional philosophy than programming. Its most important ramifications, in fact, have little to do with instructional media or methods, but more with determination of course obobjectives and with evaluation of whether the students have, in fact, achieved those objectives.

These new concepts were originally developed in a context of training for jet engine mechanics, supply clerks, and cryptographic technicians, so I would first like to describe how the concepts were applied in those courses. This will be relatively easy to do. Then I will discuss the more difficult task of applying a few of the concepts of instructional technology to the problem of teaching English as a foreign language.

It has long been customary to set training objectives on the basis of faculty estimates of what the student ought to know. Now, however, industrial and military curriculum designers are placing less and less reliance on the judgment of school staffs, since master instructors too often want to include everything they have learned in twenty years of schooling, experience, and reading.

The present trend is toward making a careful on-the-spot analysis of what mechanics or supply clerks must actually *do* to perform adequately on the job. From this inventory of observed behaviors, the instructional technologist writes a set of training objectives.

It is almost invariably found that while this list of objectives is longer because of its detail, it represents a smaller training problem than the one written up on the basis of faculty judgment. This is because the vague, the abstract, and the presumed nice-to-know items are eliminated and the course is not inflated by the ego-involvement of the experienced expert.

Now, once the instructional technologist has, from observation, determined the actual behaviors necessary for adequate job performance, he begins devising a test which will tell him, with similar objectivity, whether or not a student is able to perform them. And since his inventory of the job presumably contains a description of every necessary behavior and contains nothing that is irrelevant to adequate performance, it is only logical to assume that *every* graduate of the school needs to be able to perform *every* behavior on the inventory before he can

Mr. Cartier is Chief of the Development Division, Defense Language Institute, English Language School, Lackland Air Force Base, Texas. He is the author of *The Phonetic Alphabet* (William C. Brown Company, 1054), articles on phonetics, communication theory, and programmed instruction, and has published previously in *TESOL Quarterly* (September, 1967).

be considered ready to be assigned to the job.

Note that the instructional technologist is not interested in how well one student compares with the class mean score (the norm) at graduation, but solely in whether each individual student can demonstrate the ability to perform each and every one of the essential job behaviors (the criteria). The instructional technologist therefore speaks of his tests as being "criterion referenced" rather than "norm referenced." Students are differentiated from each other only by the amount of instruction they need in order to pass. When the amount of instruction needed becomes so great as to be uneconomical, the student is failed.

One of the most unusual aspects of this procedure is that the instructional technologist starts building his curriculum by preparing the final examination. He then builds a course that teaches the student to pass the examination. Such a procedure would be sheer insanity except for two facts. First, the test does not merely sample parts of the course, but covers *everything* the student must learn to do. Second, *every* student is expected to get *every* item right. Impossible? Not at all, though it is very difficult. However, such a procedure gives one the immeasurable advantage of being able to say to the organization that one's graduate goes to, "This man may not know everything there is to know about the job we have trained him for, but here is a list of things that we guarantee you he can accomplish, and accomplish according to the technical specifications of the job."

Now let's take a closer look at the kind of test the instructional technologist uses that permits him to make that kind of guarantee. He calls it a *criterion test*. The best way to describe it is to contrast it with the traditional kind of norm-referenced test. (Each kind has its advantages, but in the interest of brevity, I will not discuss the advantages of the norm-referenced test.) There are eight points of contrast.

1. The traditional norm-referenced test is designed to produce a normal distribution of student scores. The criterion test, however, is not designed to produce even a range of scores. A distribution is not needed since students' scores are not compared with each other.

2. A norm-referenced test usually only samples the course objectives; it is *hoped* that the student knows more than he is tested on. A criterion test tests every essential behavior.

3. Norm-referenced tests are usually satisfied with indirect testing. That is, a printed multiple-choice test with an IBM answer sheet might be used to test what the student knows about repairing an engine. Insofar as possible, a criterion test requires the student to demonstrate the actual repair procedures.

4. A student can pass a norm-referenced test even though he misses a certain pre-determined number of items. Sometimes the passing score is even determined *after* the test has been given. The number of items the student can miss and still graduate is often as high as fifty percent. On a criterion test, each student is expected to get *all* the items right, though for practical reasons we often lower that to ninety percent.

5. In grading a norm-referenced test, one does not attempt to identify *which* items a student missed; one only counts them. So one never knows what misconceptions the graduate may take away with him. The concept of criterion testing requires that each student be given at least *some* remedial training on any item he missed, even if he got the passing ninety percent.

6. For obvious reasons, test security is a constant problem with the sampling-type, competitive, norm-referenced test. But since criterion tests actually test for on-the-job competence, the student can be given full information about the nature of the test at the very beginning of the course. Indeed, the ideal criterion test constitutes a statement of the course objectives.

7. Criterion tests are much more difficult to devise and administer, but the additional time and effort is easily justified by the reliability and validity of the information they provide about student ability.

8. The last point of contrast is perhaps the most important one. If an item on a norm-referenced test is missed by a great number of students, the item is revised. If an item on a criterion test is missed by a great number of students, the *course* is revised.

Obviously, the theory of criterion testing can be applied much more readily to training for simple, mechanical jobs than to the kind of training we do at the Defense Language Institute's English Language School—teaching foreign military personnel enough English to permit them to attend technical military courses in the United States. The application of criterion testing to language training is, in fact, limited by three important factors. First, criterion testing assumes that a comple.e and unambiguous inventory can be made of all the behaviors necessary for adequate performance. Linguistic science is not yet sufficiently advanced to provide us with such an unambiguous inventory. Second, an inventory of only the most obviously essential English structures, term, and so forth needed to pursue technical military training results in several thousand individual behavioral objectives. A final criterion test with an item for each abjective would be impractically long. Third, there are no empirically-determined standards of intelligibility, of syntactic accuracy, or of many other aspects of the language, which can be applied dogmatically to assessment of a student's capability of performing the duties assigned to him after he leaves the English Language School. We must still rely on subjective judgments of pronunciation, fluency, and so on. Furthermore, these judgments are made by the wrong people; they are made by sophisticated language instructors who have become quite skilled at understanding heavily dialectal English, rather than by the student's eventual instructors, classmates, and job supervisors.

Nevertheless, it is possible to apply the theory of criterion testing to a few very important aspects of English-language training, especially since, at the English Language School, we have one enormous advantage that most schools do not have. We know exactly where the student will go, what he will be studying there, and what kind of work he will be doing afterward. Also, our job—our "mission," as we say in the armed forces—is very clearly stated. It is to turn out a student who speaks English. What do we mean by that? We mean a student who can sit down beside an American student in a classroom and learn the same things the American is being taught. We have to teach what is essential, but economy dictates that we waste no time teaching non-essential knowledge or skills.

It has therefore been necessary (and, fortunately, our circumstances make it possible) to make an empirical study of the English used by a fairly broad sample of technical-cource instructors and prepare a frequency rank distribution of the vocabulary. Like many other such lists, it shows that 93 percent of the vocabulary used is accounted for by about 1,700 words. The first few words rank much the same as in other lists. The first ten are: *the, of, and, to, a* and *an, is, in, that,* and *it.* These account for 26 percent of the vocabulary. (These same words account for 25 percent of the vocabulary in the study by Godfrey Dewey.) However, some differences show up as high on the list as the 43rd, 44th and 45th words, which are *hundred, engine,* and *pressure.* By adding some relatively infrequent but important words such as *caution, exit,* and *payroll,* we have come up with a list of about 2,300 words which will in time become the "core" vocabulary of our general English course. In addition, we have compiled similar "core" vocabularies for each of the technical specialties that our students will study when they leave the English Language School. These lists average a couple of hundred words. It is our intention to test all these "core" words with criterion tests. I hasten to add that we hope to teach more than these words, but that we will continue to evaluate those additional objectives with traditional achievement tests. We will also attempt to set core objectives with regard to English structures and other aspects, but we are putting that off until we learn to cope with the much simpler problem of vocabulary alone.

Application of this philosophy results in several deviations from the traditional methods of teaching a foreign language that you and I were subjected to in college. Since we are concerned exclusively with what the student can *do* at the end of the course, we are very little concerned with what he knows *about* the language and have eliminated all but a very few grammatical terms.

Similarly, because we find that our graduates have far less need to write English than to read, hear, and speak it, we have reduced written assignments to a minimum in order to concentrate heavily on conversation and reading.

In general, then, the school drastically limits its objectives and then singles out those which, from statistical studies or direct observation, appear to be of greatest operational value. These high-value objectives will eventually be taught to criterion. We are gradually revising our curriculum in this direction. Since we use nearly 50 different volumes and some 600 different laboratory tapes, this will take a little time.

Now let me give you some idea of what a criterion test is like. Since a criterion test is supposed to elicit the actual language behavior called for by an objective, multiple-choice items are used but rarely. Marking *a*, *b*, *c*, or *d* on an IBM sheet is not a language behavior. Multiple-choice items can test for discrimination and reading comprehension, of course, but we cannot justifiably use them to evaluate a student's ability to *produce* a word or phrase. Another objection to multiple-choice tests is the guessing factor, though the probabilities for passing by guessing are quite small when you set ninety percent as the passing score.

The theory requires that the test environment and circumstances approximate those of the work situation, which, for our students, may be a technical school, a maintenance hangar, an aircraft at 40,000 feet, and sometimes even somewhere ten fathoms deep. Those circumstances are pretty hard to duplicate, but it may be possible to set up situations in which the student must understand and respond in English under distractions and psychological pressure. And, of course, whenever the objective is comprehension of English speech, the item must be tape recorded. In fact, the great bulk of our tests have been presented aurally since long before criterion testing was ever heard of.

The new theory is forcing us to rethink the wording of individual test items, too. An item such as, "What is the meaning of the word 'hammer'?" which asks the student to think about the language, is now rewritten to read, "What do you use a hammer for?" The response might be the same in both cases, but the psychological set of the student is very different. The theory asks for more than this, though. It asks that the stem of the item be an approximation of the job situation. So another item might read, "You need to drive a nail. What do you ask for?" Note that this item calls for the student to *respond* with *hammer* rather than respond *to* the word *hammer*. This item is not, therefore, interchangeable with the others. We cannot be certain that the student who recognizes a word can use it, or *vice versa;* both kinds of items are necessary.

The theory of criterion testing increases one's sensitivity to many of the common unstated assumptions about language testing. To give just one example, the assumption that an item should consist of a language stimulus followed by a language response is implicit in most tests. This would be valid only if we made a lot of other assumptions—for example, that the students were never expected to *initiate* communication. Analysis of the actual job requirements shows that it is necessary to teach—and therefore test for—ability to make a language response to a situational stimulus, and also to respond to a language stimulus with some meaningful action other than language. So, for example, a criterion test might have items such as, "Convert the angle on your answer sheet to a triangle." Or, "What is the average of 1, 3, and 8? Write your answer in the semi-circle on your answer sheet." Also, many items will use pictures of things and activities.

It will be apparent from these examples that a single item often tests for several objectives. This complicates the post-test diagnosis of a student's specific deficiencies, but is necessary if we are to test all core objectives in a test of practical length.

One problem raised by the theory of

criterion testing is particularly difficult to solve in language training. Criterion tests insist on actual behavior—which in our case is largely spoken English. Such tests can, of course, be given in the language laboratory, but the time required to score spoken answers on tape becomes enormous when the instructor must listen through each individual tape for each student. This is especially true since the recorded answers are spaced out by the time required for the recorded question. Two possible solutions seem worth considering. First, having the student record his answers on a recorder equipped with a voice-operated relay which will run only when he is talking, or second, training the instructors to listen to speeded playbacks. Both of these are theoretically possible. A combination of them might make it practical to test in this manner, especially if we do not attempt to use such scoring methods for fine judgments of pronunciation or suprasegmental features, but only for grammar and vocabulary.

Considering all the problems of applying criterion testing to language training, it is tempting to simply throw up one's hands in despair and rationalize that the state of the art is not yet sufficiently advanced for us to bother with it. But the economic and pedagogical advantages of this approach to the defining of objectives and evaluating their achievement by the student are so great that the effort is surely justified. If we continue to set vague, general, idealistic objectives on the basis of guesswork or "experience" rather than on an objective, systematic appraisal of the student's real and immediate needs, and if we continue to pass the student who learns only a certain arbitrarily determined percentage of the language without regard to which aspects he has failed to learn, we shall never be quite sure what me mean when we say of our graduate, "He speaks English, too."