DOCUMENT RESUME

ED 112 681 FL 007 151

AUTHOR Pimsleur, Paul
TITLE Criterion vs. Norm-referenced Testing.
PUB DATE Sep 75
NOTE 5p.; Not available in hard copy due to marginal legibility of original document
JOURNAL CIT Language Association Bulletin; v27 n1 p21-24 Sep 1975

EDRS PRICE MF-$0.76 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS College Language Programs; *Criterion Referenced Tests; Evaluation Criteria; *Evaluation Methods; Formative Evaluation; *Language Instruction; *Language Tests; Linguistic Performance; *Norm Referenced Tests; Second Language Learning; Sociolinguistics; Summative Evaluation; Teaching Methods; Testing
IDENTIFIERS Communicative Competence; *Criterion Referenced Teaching; Norm Referenced Teaching

ABSTRACT

A norm-referenced evaluation system, which evaluates the student in comparison to his peers, is rejected in favor of a criterion-referenced system. The latter, which rates the performance of a student on an absolute standard, makes for an individualized approach. Two kinds of tests are distinguished, the formative, administered during the course of learning, and the summative, which evaluates achievement following a given period of instruction. It is suggested that formative tests be of the criterion-referenced type. A specific testing program is outlined whereby an aptitude test is administered before instruction begins. When instruction begins, testing is tied directly to the teaching process. The advantage of this is knowing immediately which items need more work. Criterion-referenced testing also de-emphasizes competition, reduces anxiety about learning, and encourages teachers to set realistic and reachable goals, thus helping the less able students. Future language instruction should place greater emphasis on the teaching and testing of communicative competence. The results of an experiment conducted at the University of Illinois to determine the effectiveness of incorporating competence training into the learning process reinforce the importance of teaching communicative competence. A suggested foreign language testing program follows the article. (CLK)

Criterion vs. Norm-referenced Testing

Paul Pimsleur, SUNY at Albany

Student performance has traditionally been evaluated on a norm-referenced, relative basis. The student receives a grade for his performance in comparison with his peers, rather than for the absolute value of that performance in and of itself. The norm-referenced system has pernicious effects. It implies that, for a student's grade to mean anything, there must be a reference group of students who attain lower grades than his. It insists that there be winners and losers. It creates non-achievers. How does it do this? By insisting that a test on which most students get high grades is not difficult enough. Measurement textbooks will tell you that, to insure a maximum spread of scores, your test items should be answered correctly by only 50 to 60% of the students, on the average. If 90% answer correctly, you must find harder items. The sad result is that almost half the class will be labeled "below-average." It is evident that a steady diet of punitive grades will have a disastrous effect on the morale of low-achieving students.

These considerations have led to renewed interest in the notion of criterion-referenced testing, in which a student's performance is rated by an absolute standard — has he or has he not mastered the objectives of this unit of study? — rather than in relation to his classmates' performance. If 95% of the class perform well, that's fine. This approach focuses on the objectives to be reached, saying nothing about how long it may take a student to reach them. It thereby fosters clarity in the statement of objectives and individualization in the methods of instruction.

The group working on Benjamin Bloom's concept of "mastery learning" has made a useful distinction between "formative" and "summative" evaluation. Achievement tests given at the end of major units are likely to be "summative." They sum up the students' achievement over a period of instruction. "Formative" tests are applied during the course of learning — and the sooner the better. Formative tests assess whether the students have mastered the material taught today or this week. They are as much a guide for the teacher as a measure of student performance. The teacher wants (or should want) to keep in constant touch with the students' progress — not so as to grade them but so as to adjust his teaching to their learning. It has been recommended, in fact, that no grades be given for formative tests. They are for quality control.

It seems clear that formative tests should be of the "criterion-referenced" variety. It also seems likely that certain "summative" tests will have to remain norm-referenced — at least as long as the present grading requirements are with us and we are obliged by the school authorities to spread the grades out in an approximately "normal-curve" pattern. The implication at present is that a teacher whose students all get high grades is not a good teacher but an easy marker. To be considered a fair marker, she must spread her grades over the spectrum from A to F, or at least A to D, regardless of the students' performance on any absolute standard. If her students are all doing well, she ups the standards until some do poorly. I have suggested in the past that an excellent case might be made for considering only A and B to be passing grades in foreign language, because foreign languages are more sequential than any other subject. A "C" in Spanish 1 is a virtual guarantee that the student will do very poorly in his Spanish 2 class. Therefore he should not go there.

I'd like to suggest a testing program for the average school, under present circumstances — that is to say, where individualization is far from complete and teachers are expected to give a spectrum of grades (See Addendum). This is not a visionary program, but a workable one. It begins before language study itself, with an aptitude test whose purpose is to give the school a means to section students in relatively homogeneous fashion, and — quite importantly — to identify unusual learners: those who may have special aptitude for languages or a special lack of aptitude.

Once instruction begins, testing is intimately incorporated in the moment-to-moment teaching process. After all, how is a teacher to know when a point has been sufficiently drilled unless she is receiving constant feedback from the students? Since the advent of programmed instruction, teaching and testing have been virtually inseparable. To show you what I mean, let's do a drill on the present perfect tense in French — the *passé composé*. I'll state a sentence in the present, and you state the same sentence in the past. For example, if I say *Je parle au téléphone*, you would say *J'ai parlé au téléphone*.

Je parle au téléphone. - - -
Robert pose une question. - - -
Marie donne la réponse. - - -

Good. So far it's a typical pattern drill. Now it will be transformed into a test. Now give yourself a check-mark for every one you get right, and an 'x' if you get it wrong.

Marie donne la réponse. - - -
Je reçois le paquet. - - -
Je l'ouvre. - - -
Ma sœur me regarde. - - -
Mais elle ne voit rien. - - -

Was that a drill or a test? I don't know and I don't really care — the point is that it provided instant feedback on how well we were doing at mastering the passé composé. As a teacher I would use this information to adjust my next lesson to students' needs. I would not want to wait until Friday's test, or even longer, to discover that more work on the passé composé was needed: I want to know this now. I would point out, incidentally, that quizzes like this add nothing to the teacher's burden. Since she does not intend to record the grade, she can merely ask for a show of hands to find out how students are doing. This can of course be done in the foreign language virtually from the first week of class.

The oral exam is a European idea which I've seen work well in American schools. The idea is to conduct a semi-formal oral exam in which each student appears before a committee of at least two teachers, talks for two or three minutes on a prepared subject, and then is questioned by one or both teachers. They may chat with him about topics of common interest, ask him to describe a picture or two, etc. The whole thing takes no more than five minutes. Yet it may spark a year's worth of motivation to practice speaking, because students know it is coming and that it cannot be prepared for by cramming the night before. It requires a minimum of organization (putting two classes together, for example, so the teachers can form a testing committee) and can produce excellent results in terms of putting teeth into the speaking objective. It would be wise for schools to institute the oral exam now, against the day when the speaking part of the Regents becomes more structured, as has been proposed, along the lines of the Ontario speaking test, which very much resembles the exam I have been describing.

Criterion-referenced tests are, as you have gathered, a help to the learnable student.

1. They de-emphasize competition among students in favor of competition with one's self.
2. They reduce test anxiety by allowing a student to take the test more than once.
3. They enhance understanding by stressing concrete and practical presentation derived from clear statements of course objectives.
4. They encourage teachers to set realistic, reachable goals; the goals are adjusted to one's own students rather than to a "norm" or "reference" group.

So much for the present state of affairs. Now I'd like to look ahead and see what kind of foreign language tests may be in store for us in the future.

Our present-day testing methods are being criticized with more and more regularity. John Carroll, Bernard Spolsky, and Leon Jakobovits have discussed and criticized what they call "discrete-point" testing — that is to say, the testing of specific language features one by one: an item for the imperfect of first conjugation verbs, an item for the partitive, for verbs conjugated with *être*, for the subjunctive, and so on. We spend most of our class time teaching these features one by one, and so it seems reasonable to test them one by one.

But writers like TESL specialist John Oller point out that the totality of language is greater than the sum of its component parts. Language is context, emotional impact, innuendo, human relating, gesture, grimace, and so much more. These indispensable features of natural speech are entirely neglected in the "discrete-point" approach to testing. Oller wishes to reincorporate them. He urges us to replace our linguistic model with a pragmatic model which would include not only *what* is being said, but *who* is saying it to *whom*, in *which* circumstances, and *why*.

Two main routes are being proposed these days for replacing "discrete-point" tests by others that will more nearly approximate the complexity of natural communication. These two routes are "integrative" testing and "competence" testing.

Of the two, the first can be disposed of in a few words as it is really nothing new. "Integrative" testing merely means that one should have the student perform some tasks that require him to use several language skills at once and to display now this, now that point of knowledge of the foreign language. Oller offers as examples the Cloze test and the dictation. The Cloze test is a sort of partial dictation: it is a passage in which certain words have been omitted. The student listens to the passage on a tape recording and attempts to fill in as many of the missing words as he can. Let me remark, parenthetically, that he is helped in this by his knowledge of the conditions [illegible] of the language — what words are likely

to occur where. Oller points out that we have neglected to teach what he calls the "grammar of expectancy," in which we would systematically lead our students to an ability to use all the contextual clues, both semantic and syntactic, that a native speaker uses automatically to help him understand what he hears or reads. A discussion of this would lead us too far afield, but I would like to point out that one senses, in the books and articles appearing these days, a return to the dictation technique that had been played down by the audio-lingual method. The new emphasis, however, is upon dictation as a quick and ready means to check on listening comprehension, rather than to dwell on details of the writing system.

I would like to recount an experiment in teaching and testing communicative competence which took place at the University of Illinois under the direction of Sandra Savignon. Communicative competence means the ability to function in a dynamic exchange. Linguistic competence – using foreign language words in an acceptable order – is only part of it. Communicative competence asks how successful a student would be at getting directions to the nearest pharmacy. How accurate an account could be given of an accident he had just witnessed? Could he make introductions at a dinner party?

We all know that mastery of the separate features of a language does not necessarily add up to the ability to use it for communication. But the audio-lingual precepts under which most of us operate make us cautious about introducing free communication too early. We feel the student must first acquire a basic set of structures and vocabulary, and some pronunciation habits. The assumption says that a student who tries to speak freely before he knows much of the language will make many incorrect utterances which he will later find difficult to unlearn. He may also become discouraged at his inability to speak and give up trying. Hence, we confine early activities to memorization, drilling, and "directed communication" under the teacher's strict control – "Ask Suzy what color dress she is wearing today," and so on. Our tests are largely of the discrete point variety and reflect these basic premises.

To test the validity of these beliefs, Sandra Savignon set up three classes of beginning French students – two experimental classes and one control class. The numbers were small – 12, 15, and 15 – and this is a limitation, but in all other respects the experiment was carefully done, and she got statistically significant results despite the small numbers. All three classes studied by the same modified audio-lingual method, using the time-honored Harris and Lévesque text. The students were pre-tested and found to be similar in language aptitude, verbal IQ, and high school grades. One teacher taught the two experimental groups and another the control group; the two teachers were similar in background, received the same orientation, and were supervised throughout the year by the same person.

All three classes went through the same course of study four days a week, but on the fifth day they did different things. The control group went to the language lab for a closely coordinated lab session, as has been customary at that university. The first experimental group, instead of going to the lab, received a series of cultural orientation sessions conducted in English; these were designed to create interest in French culture and thereby, hopefully, to increase their interest in learning the language. This group saw some excellent French films, discussed French politics, French art, French wine, and so on.

The second experimental group received training in communication designed not only to give them practice in conversing, but to free them of all inhibitions regarding the need for accuracy. They were made to feel free about "getting the meaning across" by whatever means they could muster. They went through sessions on greetings, leave-taking, information-getting, and information-giving, and gradually worked up to giving descriptions and reports. They could ask at any time how to say something in French, and could insert an English word if necessary so as not to get stuck and to keep on talking. Some of the talk must have been atrocious, but no matter – they went on, happily babbling in broken French.

At the end of one semester, all three groups were tested by trained native speakers in a four part test: Discussion, Information-getting, Reporting, and Description. The tests appear to have been carefully prepared and scored. I would say they did test "communicative competence."

The results will not surprise you. The group that had practiced communicating performed significantly better than either of the other two groups. On other tests of achievement (CEEB listening and reading, and final grades) there was no significant difference, though the "communicative" group had higher scores. The loss of an hour in the language lab did not affect performance in any way, which seems to indicate that the lab is useless in fostering free communication, at least as used in Illinois.

What does this small-scale experiment show? It reiterates something we know, but of which we require periodic reminding – that students learn exactly what they are taught and tested on. If our objective is communicative competence, then we need to teach and test this explicitly. We cannot assume it will grow by itself through teaching the building blocks of language. It is encouraging to note that even devoting one period per week to free communication seems to produce results. At the very least, we should think about leaving some time for free conversation on a regular basis, with no punitive correction whatever during those sessions. We also need to add a section or two to our speaking tests in which we score the student for his success in communicating rather than for linguistic accuracy. Where we will get the strength to listen to all that fractured French I do not know; but I do believe we must find it, because making mistakes – even horrendous mistakes and lots of them – is an unavoidable component of growth.

Suggested FL Testing Program

Type of test	*Crit./Norm ref.*	*When given?*	*Purpose(s)*
Aptitude Test	norm	prior to study	sectioning, guidance
Teaching/learning quizzes	crit.	almost daily	learning, teacher information
Progress tests	crit.	weekly	learning, motivation, teacher information
Review tests	crit./norm	every 4-6 weeks	motivation, grading
Midterm & final exams	crit./norm	mid-year, end of year	grading
Oral exam	crit./norm	1 or 2 per year	grading, motivate speaking
Standardized test	norm	every 2-3 years	program control, placement

Language Association Bulletin
1102 ARUSLEY ROAD
SCHENECTADY N Y 12309

ADDRESS CORRECTION REQUESTED
RETURN POSTAGE GUARANTEED

Non profit Organization
U S POSTAGE
PAID
Schenectady, N Y
Permit No. 110