DOCUMENT RESUME

ED 229 428 TM 830 329

AUTHOR Livingston, Samuel A.
TITLE Issues in Standard Setting: Some Comments, Some Suggestions, and Maybe Even a Few Answers.
PUB DATE Apr 83
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Testing; *Evaluation; Evaluation Methods; *Measurement Objectives; Program Evaluation; Psychometrics; *Standards; Testing Problems; Test Theory
IDENTIFIERS *Issues Approach; *Standard Setting

ABSTRACT

Discussed are nine questions regarding standard setting issues in educational testing: (1) Should normative or content-referenced standards be used? (2) Different standard setting methods yield different results. Does this finding present a problem? (3) Assess the adequacy of the grounding of various methods of standard setting in psychological and/or psychometric theory, (4) Should standards be validated? If so, how? (5) What are the appropriate roles of the client, the technical consultant, the test-takers, and the public? (6) To what extent should the standard setting process formally incorporate social and political considerations? (7) What are the ethical responsibilities of the technical consultant? (8) Why have developments come so slowly? and (9) What are the key short-term and long-term research problems that should be addressed? (PN)

Issues in Standard Setting:
Some comments, some suggestions, and maybe even a few answers.

Samuel A. Livingston
Educational Testing Service

A paper presented at the annual meeting
of the American Educational Research
Association, Montreal, March, 1983.

Issues in Standard Setting: Some comments, some suggestions, and maybe even a few answers.

Samuel A. Livingston
Educational Testing Service

In making up his list of nine questions for this symposium, John Meskauskas has followed an established principle of good test construction; he put the hardest questions at the end. I considered spending all my time on the easy questions at the beginning, so as not to have any time left for the hard questions at the end, but I decided not to do that. Instead, I will try to make at least one or two comments in response to each of the questions.

1. Should normative or content-referenced standards be used?

In some cases normative standards are entirely appropriate. In some cases they are not. In general, normative standards are appropriate when there is a limit on the number of people who can be placed above or below the standard. For example, you might be using a test to select students for an advanced course, and the number of places available might be much smaller than the number of students who could benefit from the course. The same thing could be true in the case of a remedial course. Or you might be setting standards in a situation where the test-takers may be competing to be recognized as outstanding. To stand out, you must be better than most of the group. The concept implies a normative standard.

I'd like to point out that an absolute standard can be based on norms information, if the norm group is not the group of test-takers who must meet the standard. If the standard is set at the 20th percentile of last year's test-takers, it is possible for 90 percent or even 100 percent of this year's test-takers to be above the standard.

Dick Jaeger, in a talk given at AERA a few years ago, made the statement that "all standards are ultimately normative." In a sense, this statement is true, since our ideas of what people should be able to do are bound to depend on what we believe people can do. But there is often a discrepancy between what a group of people can do and what somebody else thinks they should be able to do. So the distinction between normative and content-referenced standards is meaningful and important.

I'd also like to point out that the choice between the two types of standard is not always a straight "either-or" proposition. Sometimes a compromise is possible. A man from the Netherlands named Cees Beuk has devised a method for making this compromise in a systematic way. He has the judges make judgments about the passing score and about the percentage of the test takers who should pass. Of course, there is usually a discrepancy. He resolves the discrepancy by adjusting both the passing score and the pass rate until they are consistent with each other. The size of the adjustments depends on the agreement between the judges. Where the agreement is better, he makes a smaller adjustment; where the agreement is poorer, he makes a larger

adjustment. If the judges tend to agree on the passing score and disagree on the pass rate, he makes a smaller adjustment to the passing score and a larger adjustment to the pass rate. And vice versa. His paper will be published soon in the Journal of Educational Measurement, and I urge you to read it if you are interested in this issue.

2. Different standard-setting methods yield different results. Does this finding present a problem?

Yes.

What we have is a situation like the one in the saying: "A man with one watch knows what time it is. A man with two watches is never sure."

The question is which method to believe. In deciding which methods to consider and which to avoid, it's a good idea to remember the saying, "Garbage in, garbage out." Every method depends on some kind of judgment. The standard that comes out of the method is no better than the judgments that go into it. Your choice of a method should depend on what kind of judgments are likely to be most meaningful in your particular situation. For this reason, I tend to favor methods based on judgments of samples of students' performance. Unfortunately, these methods are not always practical, especially when we are testing knowledge, rather than skills. So we often have to fall back on normative methods or on methods based on judgments about test questions, such as Nedelsky's method or Angoff's method.

3. Assess the adequacy of the grounding of various methods of standard-setting in psychological and/or psychometric theory.

Standard setting is an example of decision making in the presence of uncertainty. We make decisions about students on the basis of their test scores. But, for a particular individual student, we cannot be sure our decision is correct. That is why I favor methods that are based on statistical decision theory. However, these methods require us to specify observable outcomes that we can use to classify individual students as belonging above or below the standard -- as adequate or inadequate in the knowledge or skills the test measures. Then we have to estimate the probabilities of these outcomes, as a function of the student's test score. In many standard-setting situations we cannot meet these requirements. As a result we often have to use methods that lack a firm theoretical foundation. But then we can no longer say that what we are doing is, in some theoretical sense, the correct thing to do.

4. Should standards be validated? If so, how?

What does it mean to "validate" a standard? If we had a valid criterion measure, we would use it to set the standard in the first place.

Maybe this question really means, "Should we always make sure that the standard depends in some way on the test scores of real, live

test-takers?" If it does, then I would say yes. If this kind of data is not used in setting the standard, it should be used as a "reality check" before the standard is put into effect. Otherwise, the results may be preposterous, or disastrous, or both.

5. What are the appropriate roles of the client, the technical consultant, the test-takers, and the public?

I'd like to focus on the role of the technical consultant and, to some extent, the role of the client. The role of the technical consultant is to provide technical advice and assistance. The advice should be of a type that will help the client make the important decisions: what method to use, who the judges should be, and so on. These decisions should be made by the person who has the legal responsibility for setting the standard, and that usually means the client. As the technical consultant, you can help the client make informed decisions on these issues, but as soon as you start making the decisions yourself, you have gone beyond the role of a technical consultant.

6. To what extent should the standard-setting process formally incorporate social and political considerations?

Social and political considerations are unavoidable. They are an integral part of the process, whether you make them explicit or not. When you select judges, you are choosing the people whose individual standards will be incorporated into the standard you set; you are

making a political decision. When you tell the judges what you mean by "adequate" and "inadequate" or by "minimally competent", you are making a political statement. You cannot avoid political considerations; you can only avoid making them explicit. In general, I think you should make these considerations explicit unless you have a very good reason not to. For example, you may have good political reasons for setting a limit on the failure rate. If so, then do it. But do it explicitly and openly. Don't try to bias the judgments toward a lower standard.

I once read a report of a standard-setting study that used a modification of Angoff's method. The modification was the use of a multiple-choice format for the judgments (of the percentage of minimally competent examinees who would get each question right). The multiple-choice options ranged from 10 percent to 75 percent. Now suppose you are a judge in this study. You look at a question and you decide that 95 percent of the minimally competent examinees would get the question right. If the options only go up to 75 percent correct, you're going to have a hard time expressing your opinion. What was happening here was that the person conducting the study was trying to bias the judgments toward a lower standard.

7. What are the ethical responsibilities of the technical consultant?

The consultant is responsible for making sure that the client makes informed decisions. The consultant is not responsible for making

sure the client doesn't do anything stupid. The consultant is responsible for making sure the client doesn't do something stupid without realizing the implications of it. Of course, as the consultant, you cannot take on this responsibility unless the client gives you the chance to review the standard-setting procedure -- after the plans have been made but before it is too late to make changes.

8. Why have developments come so slowly?

One reasons is that many of the important questions about standard setting tend to have the same answer: "It depends on the situation." Even when the questions are stated in such a way that they can be answered by empirical research, the answers turn out to be different in different situations. We are left with "findings" such as: "The Nedelsky method produces lower standards than the Angoff method, except when it doesn't."

Another reason that developments have come slowly is that it takes time to acquire experience, to realize our own mistakes, and to figure out how to correct them. About seven years ago, Michael Zieky and I co-authored a booklet on standard setting for a particular battery of tests. In that booklet I wrote a few things that I now realize were misleading, and one or two that were just plain wrong. Five years later we got the chance to write another booklet, which gave me the chance to correct those errors. I wish I could recall all the copies of the first booklet and give out copies of the second booklet in

exchange -- even though I still agree with most of what we wrote in the first booklet.

9. What are the key short-term and long-term research problems that should be addressed?

For several years I have thought that we needed to do research comparing different methods of standard setting. But now that we have been doing research of that type, we are finding that the results often don't generalize. In my own research I've found that the results sometimes don't even generalize from one group of eighth-grade English teachers to another.

The problem is that there are too many variables that matter. Not only are there many, many kinds of tests and test-takers and possible judges; there are also many variables involved in the implementation of each method. I am referring to such variables as the instructions to the judges, the information available to the judges, the ways the judges interact with each other, and so on.

I still think it is important to find out whether we can believe what our judges tell us. Since the answer may vary from one situation to another, we will have to set priorities. We may have to focus our attention on a few kinds of tests and a few kinds of judges. For example, one high-priority combination for research might be tests of basic skills in reading and math, with classroom teachers as judges.

I also think we should explore the use of rare events as criteria. In professional licensing we are trying to screen out the person who is likely to make a serious mistake. Fortunately, serious mistakes don't happen very often. They are rare events, and their probabilities of occurring are small and, therefore, hard to estimate. But with large samples of test-takers, we might be able to estimate the probability of a serious error as a function of the person's test score. It might be necessary to aggregate data over several years of testing. In that case, the test scores would have to be equated to make them comparable. To find a relationship, it might be necessary to classify mistakes according to knowledge and skill areas and to focus our attention on tests in those specific knowledge and skill areas. A person trying to do this kind of research is likely to encounter problems of access to the data and of confidentiality. Clearly, the obstacles are formidable. But a researcher in a position to get access to this kind of data just might be able to make an important contribution to the art of setting standards.