DOCUMENT RESUME

ED 217 069 TM 820 322

AUTHOR Livingston, Samuel A.
TITLE Assumptions of Standard Setting Methods.
PUB DATE Mar 82
NOTE 7p.; Paper presented at Annual Meeting of the National Council on Measurement in Education, (New York, NY, March 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; Elementary Secondary Education; *Evaluation Criteria; *Evaluators; *Group Testing; *Individual Testing; *Methods; Scores; Test Items
IDENTIFIERS Angoff Methods; Contrasting Groups Method; Ebel Method; Jaeger Method; Nedelsky Method

ABSTRACT

In the specific methods of standard setting in testing, judgments about individuals being tested, contrasting groups being tested, and judgments about the test items are discussed. In judgments about individual test-takers, assumptions are presented based on the knowledge and skills the test is intended to measure, the test-takers' skills at the time of testing and the true opinions of the test judges. Concerning contrasting groups of test-takers a hypothetical problem in the different distributions of a representative sample of test scores from the full population is presented. The assumptions based on judgments about test items considers the groups of people who can set standards in a meaningful way. The response analysis methodologies of Nedelsky, Angoff and Edel are discussed. A table illustrates the score distribution problem. (CM)

Assumptions of Standard Setting Methods*

Samuel A. Livingston
Educational Testing Service

A standard is an answer to the question, "How much is enough?" Any answer to this question necessarily involves some kind of judgment. The different methods of setting standards for test-takers' scores involve different kinds of judgments. Some involve judgments about individual test-takers. Some involve judgments about groups of test-takers. And some involve judgments about the items on the test.

All standard setting methods assume that the persons making the judgments are qualified to do so. All the methods assume that the judgments are meaningful, at least to the persons making them. All the methods assume that the judgments are made with reference to the purpose of the test. (If the test is intended to reflect the minimum level of basic math skills for high school graduation, the standard should not reflect the level of performance we would expect from a certified public accountant.)

But what does it mean to say that the judges are qualified? The first thing it means is that the judges are people whose standards we, the public, are willing to adopt. We may want to go further and require that the judges be a representative sample of all people whose standards we would be willing to adopt. But who are these people? In what ways must the sample be representative? These are a political questions, and they require political answers. All the psychometric skill in the world will not help you decide what points of view should be represented on a panel of judges.

---

*A paper presented as part of the Symposium "Empirical Evidence and 'Expert' Judgment in Standard Setting: Underlying Assumptions" at the annual meeting of the National Council on Measurement in Education, New York, March, 1982.

But in addition to these general assumptions, the specific methods of standard setting each involve some special assumptions. I will start with the methods I believe are the strongest, in the sense that their assumptions are the easiest to meet. These are the methods based on judgments about individual test-takers - real, live people whose test scores are available.

The first assumption of these methods is that the judgments about the test-takers are based on the knowledge and skills the test is intended to measure. Beware of using classroom grades as an indication of students' knowledge or skills. Grades often reflect many other characteristics as well: punctuality, good behavior, penmanship, class participation, diligence in doing homework, and so on. Better to get a separate judgment based explicitly on the knowledge and skills the test is intended to measure. If the test measures a mechanical or artistic or linguistic skill, the judges can observe and evaluate a sample of each test-taker's performance.

A second assumption is that the judgments reflect the test-takers' skills at the time of testing. Usually this assumption presents no problem. But watch out if there is a time lag between the testing and the judging.

A third assumption is that the judgments reflect the judges' true opinions. Usually they will. But beware, for example, if the judges are teachers who suspect that their judgments of their students may somehow be used to evaluate their own effectiveness as teachers.

If you are using the constrasting groups method, there is another assumption: that the test-takers who are being judged are a representative sample of all test-takers, in a particular way. They may have a very different distribution of test scores from the full population, but the <u>conditional probability of being judged adequate, given the test-taker's test score</u> must be the same in

the sample as in the population. That is, the students with scores of 70 to 75 who are being judged must be a representative sample of all students with scores of 70 to 75, and so on. Beware: if you select students on the basis of the judgments - e.g. 100 "masters" and 100 "nonmasters" - you will get a biased sample.

The example in Table 1 shows why. This is a made-up example, but I think you will agree that it is realistic. If the shortness of the test bothers you, just let each score level represent a five-point interval on a 50-question test. Notice that the distribution of test scores for students judged as masters is exactly the same in the sample as in the population. Likewise for students judged as nonmasters. But the percentage of the students at each score level who are "masters" is very different in the sample from what it is in the population. The passing score that minimizes errors of classification is the lowest score at which more than 50 percent of the students <u>in the population</u> are "masters." In the example, this score is 6 out of 10 questions correct. But the sample of 100 "masters" and 100 "nonmasters" would lead us to choose a passing score of 9 out of 10 questions correct.

Another group of methods involve judgments about groups of students. These methods assume that the judges can make a meaningful judgment about some reference group of test-takers whose scores are known. For example, the judges may be asked what percentage of last year's test-takers were adequate in the knowledge or skill the test is intended to measure. Or they may be asked to identify a group such that the average test-taker in the group represents the lowest level of knowledge or skill that can be considered adequate. For example, some colleges will give credit for a course to students who can score at least as well on an accreditation test as the average C student did after taking the course.

Berk's method, based on comparing the scores of instructed and uninstructed students, falls into this category. It does not involve a separate judgment for each student in the sample. Instead, it assumes that the uninstructed students involved in the comparison are typical of the unqualified test-takers - typical in terms of their test scores. Similarly, it assumes that the instructed students are typical of the qualified test-takers. But even if these assumptions are met, this method will not minimize the number of classification errors unless the test-taker population contains approximately equal numbers of qualified and unqualified persons.

Finally, we come to the methods based on judgments about test questions. I will not attempt to discuss the assumptions of Jaeger's method, because it applies only to a particular type of test and purpose of testing that I have not been involved with - basic skills tests used as a requirement for high school graduation. Besides, there is someone else on this panel who is much better qualified to discuss Jaeger's method: Dr. Jaeger.

Nedelsky's, Angoff's, and Ebel's methods all require the judges to do something that is very hard to do: to describe the way a particular kind of test-taker would respond to each question on the test. This kind of judgment is not a kind that people are accustomed to making, but these standard-setting methods all assume that you can find a group of people who can make this kind of judgment in a meaningful way.

If you are selecting the judges for one of these methods, there's one group of qualified experts you can expect <u>not</u> to be typical of all persons qualified to be judges. This is the group of people who wrote and selected the questions for the test. These people choose the questions they thought were

most important. Are you willing to assume that another group of qualified experts would have chosen exactly the same questions? If not, you can expect that the test-makers, as judges, will tend to set higher passing scores on their own test than other experts would - or, for that matter, higher than they themselves would set on someone else's test.

The logic of Nedelsky's method assumes that the judges can tell which wrong answers the "D-F student" can recognize as wrong. But this assumption is not necessary for Nedelsky's method to work; a much weaker assumption will do. The judges need only make judgments that will result in the correct distribution of the number of wrong answers eliminated: so many items with all the wrong answers eliminated, so many items with all but one, and so on.

Similarly, the logic of Angoff's method assumes that the judges can state the probability that a "minimally acceptable person" would answer the question correctly. But it is really necessary to assume only that the judges will be right on the average; their judgments about each individual question need not be correct. (The same applies, in a more complex way, to Ebel's method.)

However, even these weaker assumptions seem to me to be highly questionable. That is why I think we need to do research studies comparing different kinds of methods. Do students identified as "minimally acceptable" really perform the way judges using Angoff's method say they will? Will the Nedelsky passing score really be the score level at which a real, live student is as likely to be judged acceptable as unacceptable? These studies must use the same judges to make both kinds of judgments, or else they will not be able to separate the effects of different judges from the effects of different methods My ETS colleague Michael Zieky and I are now conducting a small-scale study of this type, and by this time next year I hope to be able to tell you about the results.

Assumptions of Standard Setting Methods*

Samuel A. Livingston
Educational Testing Service

Table 1

How a sample of equal numbers of "masters" and "nonmasters" can bias the choice of a passing score (hypothetical example).

Population Data (Unknown)

Test Score	Number of Masters	Number of Nonmasters	Total	Percent Masters
10	200	0	200	1.00
9	400	50	450	.89
8	200	50	250	.80
7	140	50	190	.74
6	50	30	80	.63
5	10	20	30	.33
0-4	0	0	0	--
Total	1000	200	1200	

Suppose we sample 100 masters and 100 nonmasters:

Sample Data (Known)

Score	Number of Masters	Number of Nonmasters	Total	Percent Masters
10	20	0	20	1.00
9	40	25	65	.62
8	20	25	45	.44
7	14	25	39	.36
6	5	15	20	.25
5	1	10	11	.09
0-4	0	0	0	--
Total	100	100	200	

*Presented at the annual meeting of the National Council on Measurement in Education, New York, March, 1982.