DOCUMENT RESUME

ED 202 852 SP 018 304

AUTHOR Halpin, Gerald; Halpin, Glennelle
TITLE Standard Setting for Educational Decision Making: An Example.
PUB DATE Apr 81
NOTE 17p.; Paper presented at the Meeting of the National Council on Measurement in Education (Los Angeles, CA, April, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Ability; *Admission Criteria; College Students; *English; Higher Education; *Minimum Competencies; Preservice Teacher Education; *Standards; Teacher Certification; *Teacher Education Programs; Test Bias; *Testing

ABSTRACT

A process for setting standards that may be adapted by educators in a variety of settings for educational decision making is presented. In this process, which was used to set minimum standards in English for admission to a teacher education program at a large university, a number of trial standards were initially set utilizing a variety of methods, all of which were analyzed for fairness and feasibility for selecting effective teachers. After careful consideration of each of the trial standards and the percentage of norm group students who would be considered incompetent with each, minimum standards were recommended for both a standardized objective test and a writing sample. (Authors/JD)

STANDARD SETTING FOR EDUCATIONAL DECISION MAKING: AN EXAMPLE

Gerald Halpin
and
Glennelle Halpin
Auburn University

Paper presented at the meeting of the National Council on Measurement in Education, Los Angeles, April 19.

Abstract

Presented, through example, is a process for setting standards that may be adapted by educators in a variety of settings for use in decision making. In this process, which was used to set minimum standards in English for admission to teacher education at a large university, a number of trial standards were initially set utilizing a variety of methods which are described. After careful consideration of each of the trial standards and the percentage of norm group students who would be considered incompetent with each, minimum standards were recommended for both a standardized objective test and a writing sample.

STANDARD SETTING FOR EDUCATIONAL DECISION MAKING: AN EXAMPLE

OBJECTIVE:

The objective of this paper is to present, through example, a process for setting standards to be used in educational decision making.

INSTRUMENT:

Missouri College English Test (Callis & Johnson, 1965). This instrument is a standardized 90-item objective test measuring grammar, capitalization, punctuation, spelling, sentence structure, and paragraph organization.

METHOD AND RESULTS:

BACKGROUND

In response to new teacher certification guidelines requiring competency in written English, it was decided at the large state university where this research was conducted that all students seeking admission to teacher education must take the Missouri College English Test (Callis & Johnson, 1965) and write a 30-minute essay on some general topic. In the subsequent process of setting standards to be used in selecting for admission those students at least minimally competent in English, a number of trial standards were first set for the Missouri test utilizing a variety of standard setting methods. Missouri test scores of 173 prospective teacher education students and 83 practicing teachers were the data in this process.

A description of the various methods utilized to set the trial standards for the Missouri test follow along with percentages and numbers of norm group students that would be judged incompetent with application of the different standards that resulted. The process used for dealing with the various discrepant trial standards to arrive at the cutting

score on the Missouri test to be used in admissions decisions is delineated. The process used to set a pass-fail standard for the writing sample is also described.

STANDARD SETTING ON THE MISSOURI COLLEGE ENGLISH TEST

Arbitrarily Selected Percentile

One standard setting method applied to the Missouri test was simply a normative or relative method (Ebel, 1979). With this approach the most competent (in our case an arbitrarily selected 67%) pass and the least competent 33% fail. Applying this percentage to the norm group scores on the Missouri test resulted in a minimum raw score of 55 for passing. The obvious drawback of this approach is that it allows the selected standard to vary according to the level of English competence of the norm group used in standard setting. If the level of competence of the norm group is high, then the standard set would be so high as to eliminate or fail well-qualified examinees in subsequent groups. If the performance of the initial standard setting group is low, however, a low standard will be set and some poorly qualified students in later groups may pass.

Chance/Ideal Mean

The second method applied to setting standards for the Missouri test was also a result of Ebel's (1979) reasoning:

1. On a well-constructed standardized test, no examinee, however weak, should actually get a score less than the expected chance score on that test, but one or two should be close to the expected chance score.
2. On a well-constructed standardized test the very best examinees should get scores at or near the maximum possible score.

3. Hence, the ideal mean score on a standardized test falls at a point midway between the maximum possible score and the expected chance score.
4. A fairly good estimate of the minimum score might then be defined as the point midway between the ideal mean score and the expected chance score.

A specific application of this model requires the following:

1. Average the lowest score and the expected chance score.
2. Average the actual mean score and the ideal mean score.
3. Define the minimum passing score as a point midway between the two averages.

Use of this method with the norm group resulted in a standard for passing the Missouri test of 39.47 raw score points which would be equivalent to the ninth percentile and would result in failing 9% of the norm group examinees. A weakness in this definition of a passing score is that it still leaves a substantial element of chance in determing the passing score. The items may be more difficult or more discriminating, or less difficult or less discriminating, than the test constructors intended. Whether an examinee passes or fails the test may be determined by the items on the test rather than by his or her level of English competence.

Item Judgment Methods

The weakness of the chance/ideal mean model may be overcome by having judges perform subjective analyses of each item on the test for which standards are to be set. With this approach, individual items are inspected with the concern being how the minimally competent person would perform on the items. In the standard setting process for the Missouri test three item judgement methods were applied utilizing 15

judges. Five of the judges were university professors with training and/or experience in English or language arts, five were advanced graduate students in English education, and five were practicing teachers of English in a local high school.

One of the item judgment methods reflected some more of Ebel's (1979) thinking. With this approach the judges were asked to rate each of the 90 Missouri test items along two dimensions, relevance and difficulty, with the reference point being the beginning teacher minimally competent in English. Four categories of relevance were used: essential, important, acceptable, and questionable. Three levels of difficulty were used: easy, medium, and hard. For each judge, the number of items he or she placed in each category was multiplied by the percentage (given by Ebel) of examinees expected to answer correctly questions in the category. The resulting products were summed and divided by the total number of items on the Missouri test to yield the standard for the individual. The mean of the standards set by the individuals within a group was the group standard. The recommended standards using this method were 63.32, 62.94, and 61.56 respectively for the university faculty, the practicing teachers, and the English education doctoral students. Averaging the standards for the three groups resulted in a single raw score standard of 62.61 on the Missouri test which is equivalent to the 58th percentile. This standard would result in 58% of the norm group examinees being categorized as failing the Missouri test.

With the second item judgment method (Nedelsky, 1954) the judges were asked to identify for each item on the Missouri test the response options the beginning teacher minimally competent in English would be able to eliminate as incorrect. The score for each item then became the

reciprocal of the remaining alternatives. For instance, if on a five-alternative item, which was the case for the majority of items on the Missouri test, a judge thought that a minimally competent teacher could eliminate two of the options, then for that item the score was 1/3. The judges proceeded with each item to obtain a standard on the total set of test items which was simply the sum of the fractions obtained for each item. Averaging the standards thus set by each of the 15 judges resulted in a raw score standard of 42.63 for the Nedelsky method. This standard would result in eliminating the bottom 12% of the norm group.

The third item judgment method applied was one recommended by Angoff (1971). Here the judges were asked to give the percentage of beginning teachers minimally competent in English they thought would respond correctly to each item on the Missouri test. The sum of these percentages was the minimally acceptable score for each judge. The minimum Missouri score with this procedure, when averaging across all 15 raters, was 56.02. Utilization of this raw score standard would result in the failure of 36% of the norm group.

If a standard were chosen utilizing an average of the Ebel, Nedelsky, and Angoff methods, then the average of 62.61, 42.63, and 56.02 would yield a standard raw score of 53.75, which would result in a standard set at the 31st percentile. Stated differently, 31% of the norm group would fail the Missouri test if the decision were made to average these three judgmental methods to obtain a standard.

Performance of Practicing Teachers

The methods presented in this section require a different type of judgment from that required to analyze items. Standard setters who favor these methods believe that judgments about human performance are usually more meaningful than judgments about test items.

Eighty-three practicing teachers representing seven schools in three districts from two southeastern states served as subjects for this aspect of the standard setting process. All 83 teachers completed the Missouri College English Test. As Popham (1978) concluded, the performance of practicing teachers is particularly informative as a sort of "reality check" to help standard setters discern whether their aspirations for pupil performance are in any sense consonant with the kinds of proficiency actually obtained in the real world by practicing teachers. In this study the mean raw score performance on the Missouri test of the 83 teachers was actually 60.52, which was about one point higher than that obtained by the norm group of university examinees and was equivalent to the median of the student group. If the average score obtained by these 83 practicing teachers were to be used as the standard, 50% of the norm group of student examinees would fail the Missouri test.

In order for additional standards to be set with these practicing teachers, the principals in each of their respective schools were asked to nominate approximately five teachers in his/her school at each of three distinct levels of competency in English. Principals were asked to identify (a) as masters those teachers whom they would allow to write an article or a report for publication which they would not need to proofread; (b) as marginal those teachers whom they would allow to write an article or report for publication, but which they would proofread; and (c) as nonmasters those teachers whom they would not allow to write an article or report for publication under any circumstance.

The quality of subsequent standard setting methods with these groups was completely dependent upon the ability of the principals to differentiate adequately among the teacher groups with regard to levels of English

proficiency that the Missouri test purports to measure. Thus, a one-way analysis of variance was computed to test for differences in the teachers' Missouri test scores. This analysis revealed that indeed the mean scores for the three groups were significantly different, F (2,79) = 10.54, $p < .001$. The large difference in average scores for the masters group ($\overline{X}$ = 68.57), the marginal group ($\overline{X}$ = 61.11), and the nonmasters group ($\overline{X}$ = 50.31) lent credibility to the principals' ability to differentiate levels of English competence.

Some might say that an ideal standard would be the mean performance of the group of practicing teachers called masters. However, the average score on the Missouri test ($\overline{X}$ = 68.57) of the 28 teachers in the masters group in this study would probably be an unrealistic performance expectation for standard setting purposes. Using such a standard would result in 78% of the university student norm group falling below this level of passing.

An alternative approach would involve applying what Livingston and Zieky (1978) referred to as the borderline group method. With this approach the mean ($\overline{X}$ = 61.11) for the marginal group of teachers would be the minimum raw score for passing the Missouri test. This standard, however, would result in 53% of the student norm group falling below the cutoff.

Although seemingly unadvisable, another approach would be to use as a standard the average performance of the group of practicing teachers labeled nonmasters by their principals ($\overline{X}$ = 50.31). However, even with 50.31 as the standard, 22% of the student norm group would fall below the cutoff on the Missouri test.

A slightly different approach to standard setting utilizing the practicing teachers that was looked at was the contrasting groups model (Berk, 1976). In this approach scores on the Missouri test for the 28

teachers in the masters group and the 26 teachers in the nonmasters group were plotted as frequency polygons. Their performance standard was based on the intersection of the two curves as shown in Figure 1.

Insert Figure 1 About Here

With this method the standard was set at 62 on the Missouri test. This standard would minimize the errors of judging as competent via the Missouri test those who would be judged incompetent by the principals or of judging as incompetent via the Missouri test those who would be judged competent by the principals. However, utilization of this standard would result in the failure of 57% of the student norm group. In order to be practically certain that admission on the basis of the Missouri test was not denied to those persons that might later by judged competent by their principals, the standard could be lowered to 42. However, this move would result in another possibly more serious error: admitting a large number of candidates who might later be judged by their principals as incompetent. Raising the standard on the Missouri test to a raw score of about 80 would virtually eliminate the number of persons who later might be judged incompetent by their principals, but this change would result in an inordinate number of admission denials to those candiates who might later be judged as competent. This change would also result in only a very small portion of persons passing the Missouri test.

The strength of the contrasting groups process is that it allows a standard to be based on an external criterion of teacher performance. The disadvantage of the model, especially in this study, is that only 28 competent and 26 incompetent teachers were involved. Most authorities

would readily agree that approximately 100 subjects in the smaller of the two groups are needed if this process is to be used. The small number of cases was also a problem herein when using the mastery group method,

Summary of Standard Setting Methods for the Missouri College English Test

Summarized in Table 1 are the results from the different standard setting methods utilized in this study. From this table it can be seen

Insert Table 1 About Here

that the various procedures would result in the setting of minimum raw score standards for passing the Missouri test that range from 39.47 to 68.57. With these respective standards, from 9% to 78% of the norm group of 173 prospective teacher education students would fail.

Recommended Standard for the Missouri College English Test

After careful consideration of the standards set by the different methods used and the percentages and numbers of persons who would fail, a committee of school of education faculty in the university where the study was conducted set a raw score minimum of 55 as the pass-fail cutting point for the Missouri test. This standard is the average of all trial standards and would be equivalent to a percentile rank for the student norm group of 34. This phase of the standard setting process involved human judgment as Glass (1978), Popham (1978), and Ebel (1979) recognized. However, due to extensive effort to base the standard on evidence from a variety of standard setting procedures utilizing quantitative data, it should be fairer than an arbitrarily selected one and it should be upheld in court should it be challenged.

STANDARD SETTING ON THE WRITING SAMPLE

Three professors who had preparation for and experience in teaching English and/or English education at the university level evaluated the writing samples (30-minute essays on a general topic) from the student norm group (N = 172). Adapting the procedure described by Coffman (1971), they chose to use the holistic method and a 10-point scale in their evaluations. After a 1-hour discussion of the rating process, the group of three raters rated seven sample essays with an average interrater reliability of .82. They rated 25 essays and again checked their interrater reliability which was .77, a coefficient they judged to be high enough for them to continue rating the final 147 papers. For the 172 essays, the average interrater reliability for the three raters was .88, which is most acceptable and a reflection of the seriousness with which the raters undertook the task.

Two other faculty members (one in English education and one in language arts) were then given a brief training session, and they subsequently categorized independently the 172 essays into two groups: competent and incompetent. They agreed that seven papers were clearly inadequate and 138 were adequate. A third judge, also a faculty member in English education, was called upon to categorize the 27 papers upon which the two judges disagreed. Altogether, these three judges categorized 152 papers as competent.

For these 152 papers, an average of the means of the ratings assigned by the three raters was computed. The obtained average of 19 was the recommended minimum standard for the writing sample. All future examinees who fail to score at least 55 on the Missouri test must have a combined rating from three judges of the writing sample of at least 19 in order to be classified as competent in written English.

EDUCATIONAL IMPORTANCE OF THE STUDY:

Educators are continuously called upon to make decisions that require the setting of standards of acceptable performance. Based on these standards some pass and some fail, some are called competent and some are called incompetent, some are admitted and some are denied admission to illustrate just three important educational decisions. In order for these decisions to be fair and to be upheld in court, they should not be arbitrary and capricious but instead should be based on carefully chosen standards. The procedures described in this paper can be adapted by educators in a variety of settings to carefully choose their standards for educational decision making.

REFERENCES

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Callis, R., & Johnson, W. Missouri College English Test. New York: Harcourt, Brace & World, 1965.

Coffman, W. E. Essay examinations. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Ebel, R. L. Essentials of educational measurement (2nd ed.). Englewood Cliffs, N. J.: Prentice-Hall, 1979.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Livingston, S. A., & Zieky, M. J. Manual for setting standards on the Basic Skills Assessment Tests. Princeton, N. J.: Educational Testing Service, 1977.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Popham, W. J. Setting performance standards. Los Angeles: Instructional Objectives Exchange, 1978.

Table 1

Missouri College English Test Standards:

Methods Used, Resulting Raw Score Standards, Percentile Equivalents, and Number of Norm Group Students[a] Failing With Each Standard

Method	Raw Score	Percentile	Number Failing
Thirty-third percentile	55.00	33	57
Average of ideal mean and chance	39.47	9	16
Ebel	62.61	58	100
Nedelsky	42.63	12	21
Angoff	56.02	36	62
Certified practicing teachers	60.52	50	87
Mastery group (teachers)	68.57	78	135
Borderline group (teachers)	61.11	53	92
Nonmastery (teachers)	50.31	22	38
Contrasting group (teachers)	62.00	57	99
Average of all methods	55.82	36	62

[a]Total number = 173.

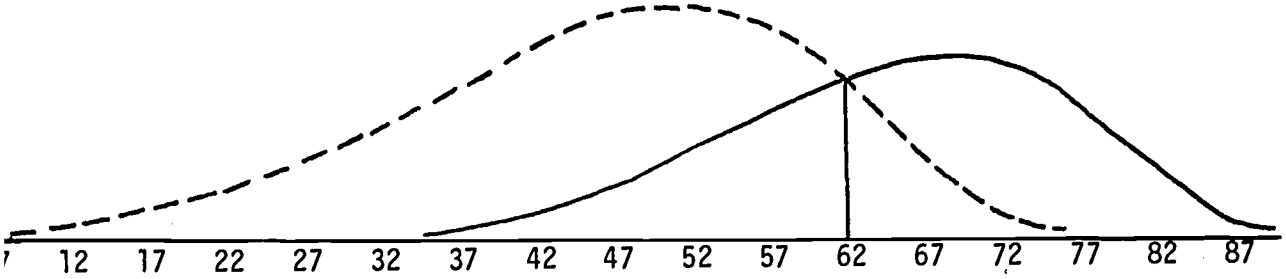

Figure 1. Frequency polygons (approximations) of Missouri test scores for masters (——) and nonmasters (- - -) teacher groups.