DOCUMENT RESUME

ED 216 020 TM 820 168

AUTHOR Poggio, John P.; And Others
TITLE An Evaluation of Contrasting-Groups Methods for Setting Standards.
SPONS AGENCY Kansas State Dept. of Education, Topeka.
PUB DATE Mar 82
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; *Basic Skills; Elementary Education; *Methods; Minimum Competencies; *Minimum Competency Testing; *Student Evaluation
IDENTIFIERS *Borderline Group Method; *Contrasting Groups Method; Kansas

ABSTRACT

Alternative group judgment approaches to setting minimum competency standards were compared. Replication of results was possible for eight different tests (reading and mathematics, across four grade levels). The Kansas Competency Based Tests in reading and mathematics were administered statewide to students in grades two, four, six, and eight. Performance standards for judging minimal competency were to be set at each grade level for each tested area. Students were rated for competence by their teachers on a four point scale. The judges' classifications of students into different competency categories defined "known groups" which provided the basis for setting performance standards. The Borderline group method and three variations of the Contrasting groups technique focus on making judgments about individual test takers. The data collected support the inconsistency of available standard setting methods in producing equivalent score standards. The authors recommend the choice of method be made with a thorough understanding of the consequences on students resulting from the level of standard set. (DWH)

Paper presented at the annual meeting of the American Educational Research Association, New York, New York, March, 1982.

An Evaluation of Contrasting-Groups Methods For Setting Standards[1]

John P. Poggio, Douglas R. Glasnapp and Dawn S. Eros
University of Kansas

With the widespread adoption of state minimum competency testing programs, the identification of an appropriate procedure for determining performance standards or passing scores becomes a major concern. A variety of procedures for setting performance standards have been proposed. Extensive descriptions of the properties and general procedures for these methods are readily available (Millman, 1973; Meskauska, 1976; Jaeger, 19.5, 1979; Glass, 1978; Hambleton, 1978; Berk, 1980; Shepard, 1979, 1980). To date, field investigations have been conducted comparing the performance of different methods within only one class of methods, those involving expert judgements of test items content (Andrew & Hecht, 1976; Jaeger, 1980; Koffler, 1980). The results from these studies are typified in a recent study by Poggio, Glasnapp and Eros (1981) that comparatively evaluated applications of the Ebel (1972), Angoff (1971) and Nedelsky (1954) approaches to setting standards. Results revealed large discrepancies in the standards produced across procedures.

A second class of procedures derive standards based on teacher judgement about the competence of the student rather than expert judgement about test item content. It has been suggested that requiring judgements about test item content may be a more contrived and difficult task than

[1] The research reported in this paper was supported by a contract from the Kansas State Department of Education.

requiring judgements about individual test takers. The inability of past research to identify a superior procedure within the former class may be reflective of this problem. A number of procedures of the latter type have been recommended as alternatives. However, within this class, the evidence necessary to judge their effectiveness and utility is not yet available (Shepard, 1980).

The purpose of the present investigation was to compare alternative group-judgement approaches to setting standards. Included were the Borderline Group method (Zieky & Livingston, 1977), and three variations of the Contrasting Groups procedure. As described by Zieky and Livingston (1977), the Contrasting Groups procedure requires identification of two groups of students, competent and not competent. The variations of this procedure examined in the present study manipulated the defining of membership of students in the competent group. Within the context of a state-wide minimum competency testing program, replication of results was possible for eight different tests (two content areas, reading and mathematics, across four grade levels, 2, 4, 6 and 8). Comparisons allowed for the description of levels and patterns of discrepancies among performance standards across methods for each replication.

METHOD

In the spring of 1980, all 2nd, 4th, 6th and 8th grade students in the state of Kansas were required to take the Kansas Competency Based Tests in reading and mathematics. As part of this testing program, performance standards for judging minimal competency were to be set at each grade level for each tested area. The number of objectives (competencies) assessed in each content area were 15, 20, 20 and 20 for the four grade levels, respectively. Three test items were used to assess each competency, resulting in test lengths of 45 items at Grade 2 and 60 items at Grades 4, 6 and 8 for each content area. Each test item was prepared in a multiple-choice format with four alternatives.

Data Collection Procedures

Approximately 60 percent (198) of the state's school districts volunteered to participate in standard setting activities. Students from a random sample of 50 of the 198 volunteering districts were rated by their teachers. Judgements were made regarding the student's level of competence on the specific state objectives being assessed in a content area. A fcur-point scale was used: (1) definitely competent on all objectives, (2) competent on most objectives, (3) minimally competent on the objectives and (4) not competent on the objectives.

To collect these data, one elementary and one junior high school building in each of the 50 districts was chosen at random. In the sampled buildings, all second, fourth, sixth and eighth grade students were rated by their teacher on the degree of competency in reading and in mathematics with respect to the state minimum competency objectives. Packets of materials containing specific directions and rating forms were distributed to teachers in the buildings selected. The rating directions to the teachers indicated that they should rate a student in mathematics or reading only if they were responsible for the student's instruction in that area. A list of the state content area competencies was included with each rating form and the teacher was instructed to carefully study and review the objectives prior to making the individual student ratings. Assurance that teachers were familiar with the state objectives and used these as the basis for their judging students was documented by way of other information gathered as part of the state testing program (Poggio and Glasnapp- 1980). Ratings of students were made prior to the actual administration of the test to the students. In all, usable standard setting data were obtained from 276 teachers, providing 13,052 ratings. The number of students rated at each grade in reading and in mathematics is provided in Table 2.

Standard Setting Methods

The methods studied focus on making judgements about individual test takers. Judges' classifications of students into different competency categories serve to define "known groups" which then provide the basis for setting performance standards. The Borderline Group method focuses only on students who boarder the minimally competent designation. The Contrasting Groups technique focuses on students classified as competent and those classified as non-competent. As with the item inspection methods, the judgements are made independent of actual test performance. However, the final standard is dependent upon actual student performance, being derived either to "maximize" correct classification of students into groups to which they are judged to belong (contrasting groups) or to evenly split the classifications of borderline students into two groups.

Borderline Group Method (BG). For this method, one group of students is identified: those whose performance is on the border of that level which differentiates competent and non-competent performance. Students classified by their teachers as minimally competent on the objectives in a content area comprised the Borderline Group. Once this group is identified, the median of the actual test scores for the group serves as the performance standard for a given test. Thus, based on actual test performance, half of the students within the identified borderline group are classified as not competent, half of them as competent.

Contrasting Groups Method (CG). For the Contrasting Groups method, the general procedure requires identification of competent and not competent groups. However, within the judged competent group, students still vary widely on their degree of competency, from just minimally competent to definitely competent. The type of student included in the competent group defines the magnitude of the discrepancy expected between the two contrasting groups and will impact the standard derived. To estimate this impact, the competent group membership was manipulated to observe three variations of the Contrasting Groups procedure. Teacher ratings classified students into one of three groups within the competent range: (1) definitely competent on all objectives, (2) competent on most objectives and (3) minimally competent. The variations used to define the competent group for the present investigation were as follows:

Contrasting Groups One (CG1): Only students assigned ratings of 1.
Contrasting Groups Two (CG2): Students assigned ratings of 1 or 2.
Contrasting Groups Three (CG3): Students assigned ratings of 1, 2 or 3.

Those students who were judged as not competent (ratings of 4) served as the contrast group for all three manipulations.

Using the group membership classification and the actual test scores of these students, a statistical likelihood-ratio procedure was used to derive the raw score standard which minimized the probability of misclassification of students in each group. There are several variants in the specific statistical procedures available depending upon the population distribution shapes and relative variances of the two groups' test scores. In the present study, the data violated both the normality and equal variance assumptions making use of the non-parametric quadratic discriminant function procedures appropriate. Throughout the present investigation, the methodology detailed by Koffler (1980) was followed, setting the "costs" of false masters equal to those of false non-masters, in all situations.

RESULTS AND DISCUSSION

Table 1 provides a framework from which to view the pattern of results that emerged from the present investigation. Included are select descriptive statistics associated with each of the 8 tests that formed the basis of the study (Poggio & Glasnapp, 1980). A review of these data suggests that the tests, based on pupil performance, provide a variety of replications over which to consider the generalizability of the present findings regarding group-judgement standard setting methods.

Insert Table 1 here

Table 2 presents the sample sizes, test score means and standard deviations for students in each of the four competency categories at each grade level. Given the state objectives, teachers identified from 3 to 7 percent of the sample as not competent in reading and 1 to 10 percent as not competent in mathematics across the grade levels. The rank order of group means confirms the expected hierarchy of level of competence defined by the rating scale categories. It should be noted that score variability increases consistently from Group 1 to Group 4 at all grade levels. The greater test score variability for students rated by teachers as either minimally competent or not competent clearly illustrates that these groups are not as well defined in terms of achievement homogeneity as are students with ratings of 1 or 2.

Insert Table 2 here

Table 3 presents the standards found using each of the four approaches to derive the cut score for tests at each grade level. The data were consistent across grade levels revealing that methods failed to identify equivalent test score standards. Rather, the CG1 procedure always resulted in a score standard substantially lower than the other approaches considered. The score standards identified by BG and CG2 methods were in the same range and varied across grade levels as to which one produced the lower or higher standard. Only on two of eight occasions did these methods result in identical score values as the standard. In summary, different configurations of groups produced vastly different standards, independent of characteristics of the tests.

Insert Table 3 here

Also included in Table 3 are the proportions of students, state-wide, who would have been classified as competent using each of the computed standards. With these data in mind, the impact of the differences among methods are seen to be even more pronounced. Given the state distributions of performances, raising or lowering the standard which defines competency one score point changes the status of approximately 4 to 6 percent of the students depending on the location of the point in the distribution. When the score standards resulting from the different procedures are discrepant by more than one or two score points, the practical impact on the number of students in a state defired as competent or not competent may be as great as 35 percent.

Rosults from this investigation can be compared and evaluated with earlier findings reported by Poggio, et al (1981) that compared standards derived from the test item judgement methods proposed by Angoff (1971), Ebel (1972) and Nedelsky (1954). Table 4 presents the Angoff, Ebel and Nedelsky score standards derived for the same tests using the same teacher population. These standards are seen to be in the same wide range as those resulting from the procedures considered in this paper. Considering all these data, there is no consistent pattern across grade levels or content area as to where each of the procedures would appear in a ranked sequence of score standards.

Insert Table 4 here

In conclusion, the data support the inconsistency of available standard setting methods in producing equivalent or even near-equivalent score standards. The important practical implication from these data is that the level of the score standard is drastically affected by the composition of the group defined as competent. Variations in the rating directions as to what kind of student is to be included in the competent group will result in widely discrepant standards. We would anticipate that a more specific definition of "not competent" would only serve to increase the variability of resulting standards.

Our findings support and strengthen the evidence that points to the arbitrariness of standards which get set in practice. The choice of method to use must be made with a thorough understanding of the consequences on students resulting from the level of the standard set. Permitting the decision as to the cut score to be left to the result of a method, like those considered in this paper, is without empirical support or justification.

REFERENCES

ANDREW, B.J. & HECHT, J.T. A preliminary investigation of two procedures for examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

ANGOFF, W.H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1971.

BERK, R.A. Criterion-Referenced Measurement: The State of the Art. Baltimore: The Johns Hopkins University Press, 1980.

EBEL, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1979.

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

HAMBLETON, R.K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-289.

JAEGER, R.M. Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 1976, 18, 22-27.

JAEGER, R.M. Measurement consequences of selected standard-setting methods. In M. Bunda and J. Sanders (Eds.) Practices and Problems in Competency-Based Measurement. Washington, D.C.: National Council on Measurement in Education, 1979.

KOFFLER, S.L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 167-178.

MESKAUSKAS, J.A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research, 1976, 46, 133-158.

MILLMAN, J. Passing scores and test lengths for domain referenced measures. Review of Educational Research, 1973, 43, 205-216.

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

POGGIO, J.P. & GLASNAPP, D.R. Report of research findings The Kansas Competency Testing Program - 1980. Topeka, KS: Kansas State Department of Education, 1980.

POGGIO, J.P., GLASNAPP, D.R. & EROS, D.S. An empirical investigation of the Angoff, Ebel and Nedelsky standard setting methods. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1981.

SHEPARD, L.A. Setting standards. In M. Bunda and J. Sanders (Eds.) Practices and Problems in Competency-Based Measurement. Washington, D.C.: National Council on Measurement in Education, 1979.

SHEPARD, L.A. Technical issues in minimum competency testing. In D.C. Berlinger (Ed.) Review of Research in Education, (Vol. 8) Itasca, Ill: F. E. Peacock, 1980.

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

Table 1

Descriptive Statistics for

the Kansas Competency Tests

Area	Grade	Items	$\bar{X}$	Mdn.	S	$\bar{P}$	N
Reading	2	45	39.6	41.7	5.9	.88	31,579
Reading	4	60	48.2	50.9	9.4	.80	33,589
Reading	6	60	45.9	48.2	9.2	.77	31,060
Reading	8	60	49.5	51.6	7.7	.83	32,067
Mathematics	2	45	42.6	43.5	3.6	.95	31,284
Mathematics	4	60	49.5	52.9	9.7	.83	33,576
Mathematics	6	60	47.6	50.3	10.0	.80	31,037
Mathematics	8	60	45.9	48.7	11.1	.77	31,999

Table 2

Group Means and Standard Deviations

		Reading				Mathematics			
Grade	Competency Level Judgement	N	%	$\bar{X}$	S_x	N	%	$\bar{X}$	S_x
2	1	546	41	43.06	2.18	610	46	43.92	1.43
	2	525	40	40.41	3.95	521	40	42.67	2.56
	3	219	16	36.43	5.59	168	13	41.01	3.01
	4	38	3	29.73	7.83	18	1	38.05	4.35
	Totals	1328				1317			
4	1	372	26	54.47	3.77	462	29	54.53	5.39
	2	646	45	50.88	5.91	662	42	51.27	6.23
	3	318	22	43.59	8.53	340	22	44.71	9.51
	4	93	7	32.34	8.47	110	7	35.27	9.63
	Totals	1429				1574			
6	1	440	30	51.66	5.68	353	23	54.04	5.67
	2	610	42	47.24	6.77	670	44	49.03	6.86
	3	305	21	39.53	8.05	365	24	41.79	8.44
	4	99	7	31.12	8.97	121	8	31.09	8.65
	Totals	1454				1509			
8	1	660	32	54.68	3.30	599	25	54.81	4.56
	2	820	40	51.29	4.75	907	38	48.69	6.94
	3	506	24	45.84	7.00	617	26	40.21	8.45
	4	82	4	37.73	10.39	239	10	34.24	10.49
	Totals	2068				2362			

Table 3

Standards Resulting from Four Approaches

		Standard				Percent				Range of Difference	
Area	Grade	BG	CG3	CG2	CG1	BG	CG3	CG2	CG1	S	P
Reading	2	37	27	36	40	79	95	82	66	14	30
Reading	4	46	42	45	47	72	81	74	69	6	12
Reading	6	40	36	42	45	78	86	73	65	10	21
Reading	8	47	39	47	51	75	91	75	57	13	34
Mathematics	2	42	0	0	41	77	-	-	84	-	-
Mathematics	4	46	42	46	47	74	82	74	72	6	10
Mathematics	6	43	38	42	44	74	83	76	72	7	12
Mathematics	8	40	30	42	48	74	90	69	54	19	36

Table 4

Standards Resulting From Three Item Judgement Methods and Four Group Judgement Methods

		Standards							
Area	Level	Nedelsky	CG3	CG2	Angoff	BG	Ebel	CG1	Range of Difference
Reading	2	22	27	36	37	37	38	40	19
Reading	4	29	42	45	43	46	43	47	19
Reading	6	28	36	42	44	40	47	45	18
Reading	8	28	39	47	43	47	48	51	24
Mathematics	2	21	0	0	40	42	38	41	-
Mathematics	4	29	42	46	46	46	47	47	19
Mathematics	6	30	38	42	43	43	47	44	18
Mathematics	8	28	30	42	39	40	45	48	21