Florida Journal of Educational Research
Fall 1985, Vol. 27, No. 1, Pp. 5-18

The Effect of Examinee Pass/Fail Information on the Level of Passing Score Set by Expert Judges

Joseph C. Saunders
South Carolina Department of Education
and
Dan L. Linton
South Carolina Department of Education

ABSTRACT. Many important applications of criterion-referenced testing require that passing or cutting scores be established. One class of procedures for setting passing scores involves the judgments of subject-matter experts. This study investigated the effect on passing scores of providing judges with pass/fail data for the various possible cut-score levels. Passing scores produced by individual judges using a modified version of the Angoff (1971) procedure are compared with the same judges' passing scores as revised after considering pass/fail data. For the seven cases studied, no consistent effect was seen on the level of passing score averaged across judges. However, a consistent reduction in the variability of passing scores among judges was observed. Judges with more extreme Angoff passing scores tended to become more moderate in their judgments after considering pass/fail information.

Many important applications of criterion-referenced testing require that passing or cutting scores be established. This practice, while widespread, has proven to be controversial (Shepard, 1978). Despite the controversy, both critics and proponents of current standard-setting practices agree on two general points. First, all passing scores are arbitrary in the sense that they are ultimately based upon value judgements. Critics (e.g., Glass, 1978)

further contend that many, if not all, passing scores are not just arbitrary but capricious, that is, selected as a result of untutored opinion rather than by reasoned judgment. Advocates of performance standards naturally disagree with this assessment (Block, 1978; Popham, 1978). Second, different standard-setting methods may, and often do, produce differing passing scores when used for a given test (Andrew & Hecht, 1976; Koffler, 1980; Mills, 1984; Skakun & Kling, 1980). This frequently observed result also has been subject to varying interpretations.

The above two points, taken together, impose a restriction upon the testing practitioner. The act of choosing a standard-setting procedure or the specific manner in which a procedure is operationalized can influence the level of passing score established. If such choices are not made on the basis of reasoned judgment, the standard-setting procedure is left open to charges of capriciousness. Of course, before sound choices can be made, the practitioner must have information about the differential effects of the available alternatives.

Many procedures for setting passing scores have been developed and used (Livingston & Zieky, 1982; Meskauskas, 1976; Millman, 1973). One well-known and widely-used class of procedures can be described generally as expert judgment methods. Such procedures typically require subject-matter experts to envision a group of hypothetical "borderline" (i.e., marginal, minimally competent, or barely passing) examinees. The judges then estimate, in some manner, the probability of a correct response for each test item (item difficulty) which could be expected given the capabilities of this hypothetical group. An unweighted or weighted sum of these estimates, averaged over all judges, becomes the test passing score. Some examples of this method include the Angoff (1971), Ebel (1972), and Nedelsky (1954) procedures.

An expert judgment method can be used by itself or as a part of a more complex procedure. For example, after using one of the above methods, judges might be provided with some sort of normative information. Such a practice is referred to as a "reality check" by Livingston and Zieky (1982, p. 57). However, few data

are available to allow the practitioner to assess the potential impact of the use of normative information. This study describes the results of providing expert judges with one type of normative data, specifically, the proportions of examinees who would pass and fail at various cut-score levels in several testing situations.

Method

Instruments

Seven different instruments, each with its corresponding committee of experts, were used in this study, and they are summarized in Table 1. Four of the tests are basic skills examinations in reading and mathematics for grades six and eight. The reading tests contain 36 items and the math tests 30 items, all written in four-option, multiple-choice format. The tests are part of a large-scale testing program currently being used in South Carolina. Two other tests, also in reading and mathematics, are designed for college underclassmen. Successful completion of these examinations is a requirement for admittance to teacher training programs in South Carolina. Each of these tests contains 56 items, also in four-option, multiple-choice format. The final instrument is a 51-item observation checklist that is used to assess the classroom performance of beginning public school teachers.

Standard-Setting Process

Standards were set for all tests in this study using a multistep procedure (Saunders, Mappus, Hamm, & Blume, 1983). In all cases, standard-setting committees were chosen so that the expert judges represented college faculty, district and school administrative personnel, and classroom teachers. During the process, judges made three different, individual decisions regarding the most appropriate cutting score. Finally, the group of judges attempted to chose a consensus passing score.

Table 1

Description of Instruments and Judges

Case Number	Instrument Type	Number of Items	Examinee Type	Number Judges
1	basic skills (reading)	36	6th Graders	28
2	basic skills (math)	30	6th Graders	25
3	basic skills (reading)	36	8th Graders	29
4	basic skills (math)	30	8th Graders	25
5	basic skills (reading)	56	college underclassmen	14
6	basic skills (math)	56	college underclassmen	14
7	classroom observation	51	beginning teachers	26

Initially, judges used the Angoff procedure to estimate the probability that a hypothetical, minimally qualified examinee or group of examinees would succeed on each item. (Estimates are made to the nearest .05, except for the observational checklist, where the estimates are to the nearest .10.) Passing scores were computed for each judge by summing these estimates over all items on the instrument. Each judge was provided with the passing score computed from his or her item estimates. Judges were then instructed to consider the test holistically and determine whether their Angoff passing scores were appropriate. If not, they were to suggest revised standards. These holistic cut scores are referred to as predata judgments. Both the Angoff passing scores and the holistic passing scores (predata judgments) provide baseline data for later comparisons.

At this point, information based on previous test administrations was given to the judges on the expected proportions of examinees passing and failing at each possible cut score for the instrument. Given this information about the consequences of their respective suggested cut scores, judges were again able to adjust their passing scores if they so desired. These standards are referred to as postdata judgments. Finally, the judges' individual passing scores were used as the basis for committee discussion leading to general agreement on a single passing score.

Analysis

In this study, the comparisons of interest are of both the initial Angoff passing scores and the holistic predata judgments with the postdata judgments. The data comparing the Angoff and holistic cut scores generally support the conclusions of a previous study (Saunders, 1983) which found no consistent difference in the two sets of cutting scores. Thus, no detailed comparison of these two sets of scores will be presented here.

For each of the seven cases studied, descriptive statistics were computed for the Angoff, predata, and postdata judgments. Mean differences between the

steps were investigated via dependent t-tests. The variability of the sets of passing scores was examined for trends. Finally, the relationships among the sets of cut scores were assessed, as well as the relationships between the level of previous cut scores and the magnitude of subsequent changes, using both Pearson product-moment and Spearman rank-order correlations.

Results

Descriptive statistics for the Angoff, predata, and postdata judgments are displayed in Table 2. In general, the passing scores set after receiving pass/fail information tended to be slightly higher than either the initial Angoff cut scores or the predata judgments. This was true whether the means or the medians of the sets of passing scores were considered. However, few of these differences reported in Table 3 are statistically significant. In two of seven cases, differences between the postdata and Angoff passing scores are significant. In these cases, one difference is negative (Case 1: -0.96) and the other is positive (Case 4: 1.84). There is one significant difference between the postdata and predata cut scores (Case 6: -1.86).

The postdata judgments do show a consistent trend toward reduced variability. As can be seen in Table 2, there are no cases in which the range of the judges' passing scores increase after receiving pass/fail information. In only one instance, Case 6, does the standard deviation of the postdata cut scores increase slightly over that of the predata judgments.

The correlations between the sets of passing scores in Table 4 reveal a consistent, positive relationship. The Pearson correlations between the postdata and predata judgments range from .61 to .78, with a median value of .75. All are significantly greater than zero. In only two instances (Cases 2 and 4) are the postdata-Angoff correlations non-significant. These correlations range from .33 to .76, with a median of .56. The Spearman rank-order correlations show very similar patterns, except that the postdata-Angoff correlation in Case 4 is statistically significant.

Table 2
Descriptive Statistics for the Angoff, Predata, and Postdata Judgements

		Angoff				Predata				Postdata			
Case Number	n*	mean	sd	median	range	mean	sd	median	range	mean	sd	median	range
1	28	24.46	3.16	24.5	13	23.96	3.20	24.0	14	23.50	2.25	23.0	12
2	25	15.16	3.21	14.0	11	15.88	2.95	15.0	11	16.32	2.25	16.0	8
3	29	22.55	4.18	23.0	17	22.45	3.45	22.0	13	22.76	2.44	23.0	10
4	25	13.76	3.73	14.0	15	15.00	3.06	15.0	11	15.60	3.04	16.0	11
5	14	39.21	8.96	38.0	32	40.36	7.74	40.0	24	42.43	3.61	42.5	15
6	14	42.86	5.33	41.0	16	44.29	4.05	45.0	16	42.43	4.40	41.5	16
7	26	40.85	6.46	43.0	26	42.15	3.93	43.0	14	42.38	3.09	43.0	13

*Number of judges

Passing Score

Table 3

Mean Differences Between Sets of Passing Scores

	Postdata-Angoff				Postdata-Predata			
Case	Mean Diff.	df	t	p	Mean Diff.	df	t	p
1	-0.96	27	-2.35	.03*	-0.46	27	-1.15	0.26
2	1.16	24	1.78	.09	0.44	24	1.12	0.27
3	0.21	28	0.36	.72	0.31	28	0.66	0.51
4	1.84	24	2.31	.03*	0.60	24	1.11	0.28
5	3.21	13	1.66	.12	2.07	13	1.39	0.19
6	-0.43	13	-0.32	.75	-1.86	13	-2.35	0.04*
7	1.54	25	1.46	.16	0.23	25	0.46	0.63

*$p < .05$, two-tailed

Table 4

Cut Score Intercorrelations

Case Number	n	Postdata-Angoff		Postdata-Predata	
		Pear-son	Spear-man	Pear-son	Spear-man
1	28	.76	.73	.75	.73
2	25	.33*	.27*	.75	.76
3	29	.68	.68	.68	.68
4	25	.33*	.34	.61	
5	14		.64	.75	.71
6	14	.50	.50	.76	.71
7	26	.56	.57	.78	.73

Note: All correlations are significantly greater than zero (at the .05 level, using a one-tailed test) unless indicated by an asterisk.

In most cases, the correlations reported in Table 5 between the judges' passing scores and the magnitude of subsequent changes show a strong negative relationship. The Pearson correlations between the judges' Angoff passing scores and their total changes in passing score (DIF31, the difference between postdata score and Angoff score) range from -.64 to -.92, with a median of -.76. The Pearson correlations for predata passing scores with postdata-predata differences (DIF32) range from -.24 to -.90, with a median of -.65. Only for Case 6 does this correlation fail to achieve significance. Again, the Spearman correlations appear very similar to the Pearson correlations, with a single additional instance (Case 4) of nonsignificance.

Discussion

Based on the data observed, knowledge of examinee pass/fail rates does not seem to have a consistent influence upon the average level of passing scores set by groups of expert judges. The lack of significant changes in mean scores after the presentation of the performance data suggests that a decision to use this type of information in standard setting should be based on considerations other than just the overall level of passing score. One such consideration might be the extent of agreement among judges. The observed reduction in the variability of the experts' passing scores, together with the high negative correlations between initial cut scores and the magnitudes of subsequent revisions, suggests that the extreme judges tend to shift toward more typical positions. This supports the use of normative information as a "reality check" since judges with more extreme views seem to be the ones whose judgments are affected most.

This study presents, in a descriptive manner, some results of actual standard-setting situations. Obviously, the generalizability of these results is quite limited. Additional limitations include the relatively small number of subject areas tested and the single method (the Angoff procedure) used to determine the initial passing scores. Nevertheless, these results should provide indications of the

Table 5

Cut Score - Change Correlations

Case Number	n	Angoff-DIF31		Predata-DIF32	
		Pearson	Spearman	Pearson	Spearman
1	28	-.70	-.59	-.71	-.73
2	25	-.76	-.81	-.65	-.56
3	29	-.81	-.75	-.71	-.71
4	25	-.69	-.63	-.45	-.33*
5	14	-.92	-.91	-.90	-.90
6	14	-.64	-.64	-.24*	-.46*
7	26	-.88	-.72	-.62	-.68

Note. All correlations are significantly less than zero (at the .05 level, using a two-tailed test) unless otherwise indicated by an asterisk.

possible effects on passing scores when judges are provided with examinee pass/fail data.

Given the widespread use of criterion-referenced tests, passing scores will continue to be set by one method or another. Since no single standard-setting procedure has established itself as the method of choice in all situations, testing practitioners must be able to make reasoned, informed decisions about the procedure to be used. Such decisions can only be made if adequate information is available about the characteristics of and differences among the various methods. Given the importance of decisions which are made based on these scores (e.g., admission, selection, certification, high school graduation), individual characteristics of standard-setting procedures should be well-documented.

References

Andrew, B. J. & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

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

Block, J. H. (1978). Standards and criteria: A response. Journal of Educational Measurement, 15, 291-295.

Ebel, R. L. (1972). Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall.

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

Koffler, S. L. (1980). A comparison of approaches for setting standards. Journal of Educational Measurement, 17, 167-178.

Livingston, S. A. & Zieky, M. J. (1982). Passing Scores. Princeton, NJ: Educational Testing Service.

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

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

Mills, C. N. (1984). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

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

Popham, W. J. (1978). As always, provocative. Journal of Educational Measurement, 15, 297-300.

Saunders, J. C. (1983). An investigation of the justifiability of passing scores based on expert judgment. Paper presented at the meeting of the Florida Educational Research Association, Orlando.

Saunders, J. C., Mappus, L. L., Hamm, D., & Blume, J. T. (1983). A multi-step procedure for setting standards based on expert judgment. Paper presented at the meeting of the Eastern Educational Research Association, Baltimore.

Shepard, L. A. (Ed.) (1978). Standard setting issue. Journal of Educational Measurement, 15(4).

Skakun, E. N. & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

AUTHORS

JOSEPH C. SAUNDERS, South Carolina Department of Education, 1429 Senate Street, Columbia, South Carolina 29201

DAN L. LINTON, South Carolina Department of Education