DOCUMENT RESUME

ED 249 275 TM 840 631

AUTHOR Halpin, Glennelle; Halpin, Gerald

TITLE Reliability and Validity of 10 Different Standard Setting Procedures.

PUB DATE Aug 83

NOTE 12p.; Paper presented at the Annual Meeting of the American Psychological Association (91st, Anaheim, CA, August 26-30, 1983).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Adults; *Comparative Analysis; *Cutting Scores; Language Arts; *Reliability; Teachers; *Validity

IDENTIFIERS Angoff Methods; Ebel Method; Nedelsky Method; *Standard Setting

ABSTRACT

Research indicating that different cut-off points result from the use of different standard-setting techniques leaves decision makers with a disturbing dilemma: Which standard-setting method is best? This investigation of the reliability and validity of 10 different standard-setting approaches was designed to provide information that might help answer that question. The 10 procedures for setting a standard on the Missouri College English Test included: a normative method (33rd percentile), the chance/ideal mean approach, the Ebel method, the Nedelsky method, the Angoff method, and five methods comparing different subsets of practicing teachers. Phi coefficients correlating pass/fail decisions for all two-method combinations of 10 standard-setting procedures ranged from .16 to 1.00 indicating greater consistency or agreement between some methods (e.g., practitioners--borderline group) than others (e.g., chance/ideal mean--masters). Phi coefficients between pass/fail with the 10 standard-setting methods and pass/fail on an external criterion ranged from .20 to .40 indicating greater validity for some methods (e.g., practitioners and borderline group) than for others (e.g., non-masters). (Author/BW)

Reliability and Validity of 10 Different Standard Setting Procedures

Glennelle Halpin and Gerald Halpin

Auburn University

Paper presented at the meeting of the American Psychological Association, Anaheim, California, August 1983

ABSTRACT

Reliability and Validity of 10 Different Standard Setting Procedures

Glennelle Halpin and Gerald Halpin

Auburn University

Research indicating that different cut-off points result from the use of different standard-setting techniques leaves decision makers with a disturbing dilemma: Which standard-setting method is best? This investigation of the reliability and validity of 10 different standard-setting approaches was designed to provide information that might help answer that question. Phi coefficients correlating pass/fail decisons for all two-method combinations of 10 standard-setting procedures ranged from .16 to 1.00 indicating greater consistency or agreement between some methods (e.g., practitioners--borderline group) than others (e.g., chance/ideal mean--masters). Phi coefficients between pass/fail with the 10 standard-setting methods and pass/fail on an external criterion ranged from .20 to .40 indicating greater validity for some methods (e.g., practitioners and borderline group) than for others (e.g., nonmasters).

Reliability and Validity of 10 Different Standard Setting Procedures

STATEMENT OF THE PROBLEM

As reported in a number of research studies (cf. Andrew and Hecht, 1976; Halpin, Sigmon, and Halpin, 1983; Koffler, 1980; Skakun and Kling, 1980), different cut-off points generally result when different standard-setting methods are used. Lacking is research which decision makers can turn to for help in choosing the best of the divergent methods. This study was designed to help fill that void with the objective being to investigate the reliability and validity of 10 different standard-setting procedures.

PROCEDURE

Pass/fail standards were set for the Missouri College English Test (Caliis and Johnson, 1965), a standardized 90-item objective test measuring grammar, capitalization, punctuation, spelling, sentence structure, and paragraph organization, using 10 different standard-setting methods. Missouri tests completed by 172 undergraduate education students and 83 practicing teachers were used in the process. A pass/fail cut-off was also set for a writing sample from the 172 undergraduate students. The standard-setting procedures for the Missouri test follow along with procedures used to set the cut-off for the writing sample.

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 this study an arbitrarily selected 67%) pass and the least competent (33%) fail.

Chance/Ideal Mean

The second method applied to setting a standard for the Missouri test was Ebel's (1979) chance/ideal mean approach which involved:

1. Averaging for the Missouri test the lowest score in the student sample group and the expected chance score.
2. Averaging the actual mean score for the student group and the ideal mean score (midway between the maximum possible score and the expected chance score).
3. Defining the minimum passing score as a point midway between the two averages.

Item Judgment Methods: Ebel, Nedelsky, and Angoff

With Ebel's (1979) item judgment method, 15 raters (five university professors in English or language arts, five doctoral students in English education, and five high school teachers of English) were asked to categorize according t relevance and difficulty each of the items on the Missouri test. The number of items in each category was multiplied by the percentage 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 a standard for each group of raters, which, when averaged across raters, resulted in the standard for this method.

For the Ebel method the average interjudge reliability for the 15 judges using a two-factor analysis of variance without replications was .84. The Pearson correlation coefficient between Ebel item ratings and actual item difficulty (proportion of student sample group responding correctly) was .49 ($p < .001$). This coefficient provides evidence of the validity of this approach since difficulty is an integral part of the Ebel ratings which, therefore, should correlate with actual item difficulty.

With the Nedelsky (1954) method, the 15 raters 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. The sum of the fractions so obtained became the standard for each rater. Averaging the standards for the 15 raters resulted in the standard for this method.

The average interjudge reliability for the Nedelsky procedure was .74. Ratings using the Nedelsky procedure indirectly entail judgments regarding item difficulty and should therefore, if valid, correlate with actual item difficulty. The obtained Pearson correlation coefficient was .24, a significant ($p < .05$) although not impressively high value.

With the Angoff (1971) method, the 15 raters were asked to give the percentages 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's minimum score.

The average interjudge reliability for the Angoff method was .81. As is true with the Nedelsky method, item judgments using the Angoff procedure require judgments of item difficulty and also should correlate with obtained item difficulty. The Pearson correlation coefficient was .57 ($p < .001$).

Performance of Practicing Teachers

Eighty-three practicing teachers representing seven schools in three districts served as subjects for this aspect of the standard setting process.

All 83 teachers completed the Missouri College English Test. The mean performance of these teachers was used as the practicing teachers standard.

In order to set additional standards with the practicing teachers, the principals in each of the seven schools were asked to nominate five teachers at each of three distinct levels of competency: masters, marginal, and nonmasters. The mean on the Missouri test for the masters group was the master teachers standard. Applying what Livingston and Zieky (1978) referred to as the borderline group method and using as the standard the mean of the marginal group resulted in the borderline teachers standard. The mean of the nonmasters group became the nonmaster teachers standard.

The final approach to standard setting utilizing the practicing teachers was the contrasting groups model (Berk, 1976). With 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. The standard was then set at the intersection point of the two curves.

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 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 on 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 was the recommended minimum standard for the writing sample.

ANALYSES, RESULTS, AND CONCLUSIONS

To get an indication of the reliability (equivalence) among the standards set for the Missouri test using the 10 different methods, Phi coefficients were computed using pass-fail decisions for all possible two-method combinations. To investigate the validity of the standard set with each of the 10 methods, Phi coefficients were computed correlating pass/fail decisions on the writing sample with pass/fail decisions based on each of the 10 standard setting methods.

As shown in Table 1, the resulting Phi coefficients for all the two-method combinations of the 10 approaches to standard setting ranged from .16 to 1.00 with the median being .53. The least reliable or consistent methods in this study were the chance/ideal mean--master teachers and the Nedelsky-master teachers approaches. Based on these results, standards set with either of these two-method combinations are likely to differ. The most reliable or consistent were the practicing teachers--borderline teachers, Ebel-contrasting groups, and 33rd percentile-Angoff methods. These findings indicate that similar standards are likely to result from the use of any of these three two-method combinations.

Also as reported in Table 1, the resulting Phi coefficients between pass/fail decisions on the writing sample and pass/fail with each of the 10 approaches to standard setting ranged from .20 for the nonmastery method to .40 for the practicing teachers and the borderline teachers cut-offs with the median being .31. Thus, some methods (e.g., Nedelsky and chance/ideal mean) do appear to be less valid than others (e.g., practitioners and borderline group), at least for setting standards on the Missouri College English Test that correlate with performance on a writing sample.

Psychologists in a variety of specializations are often 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 decisions. In order for these decisions to be fair and to be upheld in court, they need to be based on reliable and valid standards. Results of this study provide much-needed information about both the reliability and the validity of 10 different standard-setting procedures which should be useful to all faced with the task of choosing standards for decision making.

References

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

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 and 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, New Jersey: Prentice-Hall, 1979.

Halpin, G., Sigmon, G., & Halpin, G. Minimum competency standards set by thre 'vergent groups of raters using three judgmental procedures: Implications for validity. *Educational and Psychological Measurement*, 1983, 43, 185-196.

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

Livingston, S. A., & Zieky, M. J. *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, N.J.: Educational Testing Service, 1978.

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

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

Table 1

Intercorrelations Among Pass-Fail Decisions Using 10 Different Standard Setting Methods and an External Criterion

Standard Setting Methods	Phi Coefficients										
	1	2	3	4	5	6	7	8	9	10	11
1. 33rd Percentile		.45	.57	.53	.96	.69	.36	.69	.73	.59	.32
2. Chance/Ideal Mean			.26	.84	.43	.31	.16	.31	.61	.26	.22
3. Ebel				.30	.60	.83	.62	.83	.42	.97	.33
4. Nedelsky					.51	.36	.19	.36	.72	.31	.21
5. Angoff						.72	.37	.72	.70	.62	.31
6. Practicing Teachers							.51	1.00	.50	.86	.40
7. Master Teachers								.51	.26	.60	.25
8. Borderline Teachers									.50	.86	.40
9. Nonmaster Teachers										.43	.20
10. Contrasting Groups											.32
External Criterion:											
11. Writing Sample											

Note: For Phi coefficients .16 and .19, $p < .05$.
For Phi coefficients .20, .21, .22, and .25, $p < .01$.
For Phi coefficients ≥ .26, $p < .001$.