DOCUMENT RESUME

ED 205 552 TM 810 431

AUTHOR Poggio, John P.: And Others
TITLE An Empirical Investigation of the Angoff, Ebel and Nedelsky Standard Setting Methods.
INSTITUTION Kansas Univ., Lawrence.
SPONS AGENCY Kansas State Dept. of Education, Topeka.
PUB DATE Apr 81
NOTE 29p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis: *Criterion Referenced Tests: *Cutting Scores: Elementary Secondary Education: Evaluation Criteria: *Minimum Competency Testing: Quantitative Tests: Reading Tests: *Scoring Formulas: Test Validity
IDENTIFIERS Angoff Methods: Contrasting Groups Method: Ebel Method: Kansas: Nedelsky Method

ABSTRACT

A comparison of four frequently used standard setting methods for deriving cut-off scores with respect to the expected performance of minimally competent students, is presented in this paper. Ten Kansas Competency Based Tests, in reading and mathematics, were administered across five grade levels in a state-wide minimal competency testing program. School districts were randomly assigned the Angoff, Ebel or Nedelsky standard setting method and 50 districts were assigned the Contrasting Groups Method. Descriptions of the types of judgments required and the procedures used for deriving a standard are given for each method. Methods of analysis are documented and are followed by results of their discussion. Statistical evidence is provided in the eight tables appended. Table 8 provides a framework within which to view the pattern of results for the ten tests. Data indicate that performance score standards are consistently ranked, and that discrepancies were substantial between methods but internal consistency was high. It is concluded that, since standard setting methods differ and competency level is a continuous variable, methods are bound to produce different results. The superiority of a single method is neither supported by existing literature nor data. (Author/AEF)

Presented at the Annual Meeting of the American Educational Research Association, Los Angeles, California, April, 1981.

An Empirical Investigation of the Angoff, Ebel and Nedelsky Standard Setting Methods[1]

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

Since the concept of criterion-referenced testing was introduced during the 1960's (Glaser, 1963; Popham & Husek, 1969) much attention has focused on criterion-referenced measures. Acceptance and application of the concept has shown tremendous growth. An example of the widespread adoption of criterion-referenced measures is reflected in the minimum competency testing movement that began in the early 1970's. Since that time, 38 states have taken some sort of formal action involving competency-based assessment. The debate on the merits and demerits of minimum competency testing continues. Regardless of the length and breadth of these debates, the reality is that minimum competency testing is occuring. An issue which arises within a competency testing program is that of determining the performance standards or passing scores.

A variety of alternative procedures for setting standards have received much attention in the literature. Extensive description of the properties of the methods, their underlying assumptions, the purposes each addresses and general procedures to follow when setting standards are numerous (Millman, 1973; Meskauskas, 1976; Jaeger, 1976; Glass, 1978; Hambleton, 1978; Hambleton, Powell & Eignor, 1979; Shepard, 1980). However, in the opinion of many of these authors, there is little empirical evidence to offer users the necessary guidance in choosing among these methods for setting standards. To date,

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

only a few investigations have been conducted comparing the performance of different methods. The results of these studies tend to be mixed. For example, Andrew and Hecht (1976) compared the procedures proposed by Nedelsky (1954) and Ebel (1972), as did Skakun and Kling (1980). The latter, however, used a modification of the Ebel procedure which prevents generalizability of the results to non-similar situations. The Nedelsky procedure was also used in studies by Brennan and Lockwood (1979) and Koffler (1980), comparing it to a procedure proposed by Angoff (1971) and a procedure known as the Contrasting Group methods, respectively. Across studies, there is a lack of consistency in type of design employed, which standard setting methods were used, whether focus was solely on the actual standards derived or whether an examination of the methods' psychometric properties were included. Hence, there is still little conclusive comparative evidence to assist users in choosing among alternative standard setting methods.

The purpose of the present investigation was to simultaneously compare four frequently used standard setting methods: Angoff, Ebel, Nedelsky and Contrasting Groups. Within the context of a state-wide minimal competency testing program, comparisons of the four derived cut-off scores were possible for ten different tests (two content areas, reading and mathematics, across five grade levels, 2, 4, 6, 8 and 11). However, comparisons were not limited to the description and pattern of discrepancies among standards across methods. Rather, evaluation of the procedures also included extensive examination of the psychometric properties of each, including reliabilities and validities of both judges' ratings and judgments about students using each standard.

METHOD

Context of Data Collection

During the 1979-80 school year, all 2nd, 4th, 6th, 8th and 11th grade students in the State of Kansas were required to take the Kansas Competency-Based Tests in reading and mathematics. The number of competencies assessed in each content area were 15, 20, 20, 20 and 19 for the five grade levels, respectively. Three test items were used to assess each competency resulting in test lengths of 45 items at Grade 2, 60 items at Grades 4, 6 and 8 and 57 items at Grade 11 for each content area assessed. All test items were in a multiple-choice format. Each item was presented with four alternatives.

As part of this testing program, performance standards for judging minimal competency were required for each grade level and test area. Because no one method available appeared to be superior, it was decided to collect standard setting data using a variety of procedures. A synthesis of these data were used to set the performance standards for the State.

Data Collection Procedures

Approximately 60 percent (198) of the State's school districts volunteered to participate in the standard setting activities. Each participating district was randomly assigned one of three methods (Angoff, Ebel or Nedelsky) to use in setting standards. In addition to using a specific procedure, each district was requested to provide standard setting responses for six of the ten tests available. The pattern for which of the six tests was assigned to a district was chosen from one of ten patterns to assure an approximately equal number of ratings for each test using each procedure.

Within each district, the test coordinator was directed to select an experienced educator at each grade level and content area for every one of the six different tests assigned to the district. It was suggested that the individuals who were to participate in setting standards be knowledgeable and have at least two years teaching experience in the content area at a specific test grade level. Six packets of materials, each containing a copy of the test to be rated and a set of instructions for the standard setting procedure to be used, were sent to the test coordinator for distribution to the educators selected. The raters reviewed and judged the tests, based upon the set of instructions detailing the method they were to use, just prior to the actual administration of the tests to the students. The responses of each rater were returned to the district test coordinator who forwarded them to the investigators.

In all, usable standard setting data was obtained from 926 teachers. Examination of the demographic characteristics of the group indicated that teachers selected by their districts to participate were well qualified for the task in terms of years teaching experience ($\overline{X}$ = 13), years with the district ($\overline{X}$ = 8.3), professional training (68 percent with work beyond bachelors degree) and type of responsibility (97 percent currently teaching the content at the grade level on which they set standards). Table 2 identifies the number of respondents who provided data for a given grade level by area test using a specific procedure. The group sizes ranged from a low of 24 (Nedelsky - Reading - Grade 6) to a high of 41 (Ebel - Math - Grade 11).

In addition to the judgmental ratings of test items using the Angoff, Ebel or Nedelsky procedures, data also were collected appropriate for setting standards using the Contrasting Groups Method. The test coordinators from a representative sample of 50 districts were requested to randomly select one elementary and one junior high building in their district. At these buildings, all second, fourth, sixth and eighth grade students were to be rated by their teacher on overall competence in reading and mathematics given the State competencies. Packets of materials containing specific directions to the building principals and teachers and rating forms accompanied the testing materials. The rating directions to the teacher indicated that they should only rate a student in math or reading if they were responsible for the student's instruction in that area. A list of State content area competencies were included with each rating form and the teacher was asked to review these objectives prior to making the individual ratings. All teacher ratings were recorded in a special codes section on the student's test answer sheet. Table 1 identifies the number of students rated by grade level and content area. The number of students rated by teachers as minimally competent and not minimally competent also are provided. In total, 12,575 ratings were received (6278 in reading, 6297 in math).

Insert Table 1

Standard Setting Methods

As noted by Jaeger (1979) all standard setting methods involve judgmental decision making at some level and differ only by the ". . . proximity of the judgment--determining data to the original performance." (p. 48) As such, he suggests classification of methods under either a proximal (direct) or distal (derived) model, referenced as judgmental or empirical models, respectively. Jaeger also includes the judgmental--empirical combination methods within the proximal model.

Within the context of the present study, use of methods included in the proximal model classification was appropriate. The Angoff, Ebel and Nedelsky procedures are based on expert judges' assessments of individual items content with respect to the expected performance of minimally competent students.

Although the specific manner in which items are assessed differs across the methods, all are purported to yield an overall score which would or should be attained by minimally competent students, thereby providing a standard for judging the competence of an individual. For these methods, both the judgments made and the standards derived are independent of the actual performance of students on the test. Using these procedures, passing scores may be set prior to the test administration.

The fourth method used, Contrasting Groups, also involves expert judgment. However, the focus is on making judgments about actual individual test takers rather than on test item content. Judges' classification of students as either competent or non-competent serve to define two "known groups." As with the item inspection methods, the judgments made are independent of actual test performance. However, the final standards are dependent on performance, being derived to "maximize" correct classification of students into the groups to which they're judged to belong. For each method, a brief description of the type of judgment(s) required and the procedure used for deriving a standard follows.

Angoff method. For each item, our educators were asked to estimate the probability (on a scale of 0 to 100) that a minimally competent student would correctly answer the item. In essence, the judges were estimating the difficulty level of an item referencing a hypothetical group of individuals who would be judged as minimally competent. To obtain the overall standard, probabilities assigned by a judge were converted to proportions and summed; the average of these sums, across judges, provided the final passing score. In effect, this standard represents the (estimated) mean total score for a group of minimally competent individuals. As such, any student scoring below the mean of this reference group would not be judged to be competent.

Ebel method. For this method, judges made three separate types of judgments. First, judges rated each item on two separate dimensions: level of difficulty (easy, medium or hard) and level of relevance (essential, important, acceptable or questionable). For each judge, then, all items could be classified into one of 12 cells in a 3 x 4 grid defined by the three difficulty and four relevance categories. Judges then indicated the percentage of items within each of the 12 cells that a student should answer correctly in order to be judged minimally competent. To derive the standard from these data each item was assigned to one of the twelve cells based on the teachers' ratings. The percent passing judgment for a cell was then multiplied times the number of items in a cell and these products were summed over all 12 cells to get an overall passing score for a judge. These passing scores were then averaged over judges to get the composite passing score. Unlike the Angoff and Nedelsky procedures, interpretation of the meaning of this standard is not as precise. While difficulty of an item is reflected, it is not necessarily the major determinant. Differential weighting due to item relevance and whether judges indicate that, for example, more "hard" items of some sort "should" be answered correctly than "easy" items of another sort will affect the overall level of the standard.

Nedelsky method. For each item, judges were asked to indicate which, if any, distractors the minimally competent student should be able to eliminate as incorrect. The score for an item then became the reciprocal of the number of alternatives not eliminated. For an individual student this score represents the probability of correctly answering the item, the "chance score." Although obtained through a different process, these scores carry the same type of interpretation as those from Angoff. For a (hypothetical) group of minimally competent students, the score represents the item's difficulty level. To obtain the overall standard, the item scores are summed for each judge and these sums are then averaged across judges. This standard also represents the mean total test score expected from the reference group.

The Nedelsky method also includes a component which allows the user to determine what percentage of the group of minimally competent students should fall above and below the standard. The assumption is made that the variability in individual judge's standards equals that of the total scores of the reference group. Using the standard deviation of the judges scores as an estimate of that of the reference group, adding or subtracting a constant number of standard deviation units (_k_) to the original passing score decreases or increases, respectively, the number of minimally competent students judged as competent. There does not appear a consistent recommendation in the literature for the value to be assigned to _k_. For the purpose of this study, _k_ was assigned a value of 1.

Contrasting Group method. For a sample of students who will have scores on the test, each is classified by a judge into one of two groups, Competent or Not Competent, relative to the content being assessed. Based upon this group membership classification and the actual test scores of these students, a standard is derived using statistical likelihood ratio procedures which minimize the probability of misclassification of students into groups.

There are several variants in the specific statistical procedures available. Choice of the appropriate variant is dependent upon the population distribution shapes and relative variances of the two groups' test scores. When both groups are normally distributed and have equal variances, the Linear Discriminant Function (LDF) derived by Fisher (1936) is the appropriate statistic.

With normality but unequal variances the appropriate likelihood ratio statistic is the Quadratic Discriminant Function (QDF). Finally, with non-normal distributions, non-parametric analogs to the LDF and QDF (for equal and unequal variances, respectively) are appropriate. In the present study both the normality and equal variance assumptions were violated, making the non-parametric QDF procedures appropriate. Throughout the present investigation, the methodology detailed by Koffler (1980) was followed, setting the "costs" of false positives and false negatives equal in all situations.

Methods of Analysis

Several different indices were available to serve as the units of analysis. The Angoff, Ebel and Nedelsky methods provided individual item ratings on one or more dimensions for each judge plus a total test composite standard was derived for each judge. In addition to the data provided by the judges, item difficulties and total test score distributions were known based on actual student performance data.

All analyses performed were descriptive in nature. The intent of the analyses were to provide comparative descriptive information resulting from application of each of the methods. The statistical information provided includes:

1. Distributional characteristics of the judges' composite scores within a group.

2. Indices of reliabilities of the judges' ratings.

3. Correlations among the various indices resulting from the standard setting methods using the item responses as the units of analysis.

4. Squared-error loss reliability (Brennan, 1980) estimates (indices of dependability) for total test scores given a passing score determined by a standard setting procedure.

5. Classification agreement tables on student designation of competent based on student performance data using the passing score determined by each method as one classification criteria and teacher judgments as the other criteria.

All results are provided for ten replications across the five grade levels and two content areas.

RESULTS

Table 2 provides information on the distributional characteristics of the Angoff, Ebel and Nedelsky methods. These indices were calculated using the total test composite ratings as the unit of analysis. Several points should be noted from these results. The distributions of the ratings for all methods have a consistent negative skew, although in most cases it is only slight. As expected in negatively skewed distributions, the means tends to be slightly smaller than the medians. The difference in these two measures of central tendency is negligible in most instances, particularly for the Ebel and Nedelsky procedures. Greater discrepancies exist in the Angoff procedure where ratings tend to be more negatively skewed than in the Ebel or Nedelsky methods. The last distributional descriptive index of importance is the variability of the judges' ratings. As indicated, the Angoff procedure results in ratings which are considerably more variable than for either of the other two procedures.

The most important comparative information in Table 2 is the resulting performance standard suggested by each procedure for the same test. Using the means as the performance standard, the Ebel procedure consistently results in the highest score. The Angoff procedure identifies a passing score in the same region of the distribution, but is generally one to five score points lower. Only in the case of the second grade mathematics test was the Angoff procedure considerably higher than the Ebel method. The suggested passing scores from both of these procedures were substantially higher than those resulting from the Nedelsky procedure, even if the value of k is set at 3 or 4.

Insert Table 2

As a measure of the consistency of ratings across items and judges, the reliability of ratings on several dimensions were computed using analysis of variance methods resulting in alpha (α) coefficients of internal consistency. For the Angoff procedure, only one rating was available for use in the calculation of a reliability index. The Ebel procedure provided four ratings, item difficulty, item relevance, the assigned cell percentage and the item composite based on the other three ratings. Two scores result from the Nedelsky procedure, the specific alternative(s) selected and the number of alternatives selected per item. Reliability coefficients were calculated for each of these scores, the one Angoff item rating, the four Ebel measures and the two Nedelsky scores. These reliability coefficients for each procedure are given in Table 3 for each grade level and content area tested. All of the coefficients are high with .89 the lowest coefficient. The individual Ebel ratings tend to be lowest (lower to mid .90's), but the judgment reliabilities for the item composite scores are high for all three methods (.98 and .99). The consistency of the judges in rating the items appears to be exceptionally stable for all three procedures.

Insert Table 3

Information in Table 4 presents bivariate correlations between actual item difficulties on the Kansas Competency Tests and the item probabilities assigned by judges to the items in the standard setting process using Angoff, Ebel and Nedelsky methods. In reviewing these data recall that in applying the standard setting procedures the referent group is "minimally competent students." Item difficulties were computed across all students tested at a grade level (N = 32,000). A number of patterns emerge in these data. Overall the Angoff method shows the highest levels of association with the actual item difficulties. The Ebel difficulty component closely parallels that of the Angoff appraisals. However, the Ebel difficulty ratings (easy, medium, hard) when blended with the remaining facets of the method produces a composite evaluation that is not at all consistent with the observed item difficulties. Item difficulties when correlated with the Nedelsky item probabilities tend to be lower and less variable over replications than the other methods.

Insert Table 4

To further explore these data, item probabilities were then correlated over methods. Results from this analysis are given in Table 5. Most apparent in these data are the rather high concurrent coefficients between the Ebel difficulty item evaluations and the ratings of items by the Nedelsky and Angoff methods. The pattern appears to be consistent with the exception of very low correlations at Grade 8. Correlations between the composite item probabilities tend not to be as high. However, the correlations between Angoff and Ebel item composite ratings are considerably higher than between Nedelsky and either of the other two. The pattern of correlations suggests the presence of the difficulty component affecting ratings across all three methods, yet beyond this aspect each method appears to capitalize on aspects of specific variance unique to itself.

Insert Table 5

The next stage of analyses addressed the issue of characteristics resulting from classification given the test standard/criterion. Table 6 reports indices of dependability given select criterion scores on each of the Kansas tests. The method used to produce these indices has been identified as a squared-error loss approach (Berk, 1980; Kane and Brennan, 1980). This method has two characteristics worth noting. First misclassification (master/non-master) further from the cut-score is treated more seriously than misclassification based on scores close to the cut-score. Second, the distribution of coefficients of dependability is "u shaped" and coefficients are lowest as the cut-score approaches the raw score mean. Coefficients are computed based on a single test administration. Values reported in Table 6 are not corrected for chance placements.

Insert Table 6

The values underscored in Table 6 are the agreement coefficients associated with the cut-score produced by each method investigated for each test. The minimum passing score derived from the Nedelsky procedure was treated as the mean rating assigned over judges plus one standard deviation unit of the judges ratings ($\bar{X}$ + 1$\underline{\sigma}$). Criterion scores for the remaining methods are based on the rating over judges. A consideration of these data shows the Nedelsky minimum passing score over tests resulting in the highest agreement coefficients for the methods considered. The lowest coefficients result with the Ebel method. These results are due to the fact that the standard from the Ebel procedure consistently approximates the actual test raw score mean. Overall, however, the observed indices of dependability appear higher. Although consistent discrepancies are noted, the magnitude of these discrepancies are mediated by the specific techniqe used and not correcting for chance placements.

Table 7 summarizes the results of classification based on teacher judgment and score classification derived using the Nedelsky, Ebel, Angoff and Contrasting Groups approach. Raw score frequencies are reported. The derived minimum passing score (standard) found for each procedure is also given. For this analysis teacher classification at each grade level and in the content area tested served as the criterion variable. The Contrasting Groups standards were derived from these teacher classifications (see Methods section). As such the fit of this procedure to the actual classification of obtained test scores might be expected to be quite stable for the sample of students for whom these data were available.

From the data presented a number of findings emerge. It is evident that the rank ordering of the procedures studied finds the Nedelsky method resulting consistently in the lowest raw score standard, followed by the Contrasting Groups standard, then the Angoff standard, while the highest standard is yielded by the Ebel method. The pattern of standards being computed for tests is one of substantial variability. That is, for the most part the standard suggested by methods for a given test tend to be quite disparate. For a given test, the order of the performance standards resulting from the four procedures is consistent. The Nedelsky procedure always results in the lowest standard followed by the Contrasting Groups, Angoff and Ebel in that order. The four procedures also result in performance standards which tend to have different degrees of variability in location across the ten different tests. The performance standard identified by the Contrasting Groups ranged from 50 percent (Math-Grade 8) to 70 percent (Reading and Math-Grade 4) of the items correct. This range ignores Math-Grade 2 where the resulting standard was set at a score of zero. The

Angoff procedure resulted in performance standards from 65 percent to 88 percent of the items correct. The range for the Ebel method was 72 percent to 84 percent of the items correct and for Nedelsky 47 percent to 50 percent of the items correct.

Across grade levels the number of students judged as not competent by teachers in grades 2, 4, 6 and 8 were 3 percent, 7 percent, 7 percent and 4 percent, respectively in reading and 1 percent, 7 percent, 8 percent and 9 percent, respectively in mathematics. Unfortunately, there was considerable variability and overlap in the actual test performance scores for students judged as competent and not competent by the teachers. The passing score derived from the Contrasting Groups procedure minimizes the number of misclassification errors given the student performance data. Using the teacher judgments as the criteria for correct classification, the frequency of a specific type of classification error for the other three procedures is dependent on how far their standard is below or above the Contrasting Groups' standard. The Nedelsky approach results in a greater percent of false masters, while the Angoff and Ebel standards reduce this form of error while increasing the occurence of false nonmasters. In considering these data, recall that the Nedelsky approach requires defining a value, k, which is then applied to the mean rating over judges. In practice this value would be set based on discussion and negotiation. For the present study this value was taken as 1 based on suggestions from Nedelsky's writing on the topic (1954). Had this value been different, either greater or smaller, then the classification data in Table 7 would change.

DISCUSSION

Table 8 provides a framework within which to view the pattern of results that emerge from this investigation. Presented are select descriptive statistics associated with each of the 10 tests that formed the basis of this investigation (Poggio and Glasnapp, 1980). A review of these data suggest that the tests, based on pupil performance, provide a variety of replications over which to consider the generalizability of the findings of this investigation of standard setting methods.

The performance score standards resulting from the application of the four procedures are consistently ranked in the same order with Ebel producing the highest standard, then Angoff, then Contrasting Groups and Nedelsky with the lowest standard. The result that none of the procedures consistently produce the same standard support the findings from previous studies (see, for example, Andrew & Hecht, 1976; Koffler, 1980; Skakun & Kling, 1980). The added significance of the present finding is that no previous study had compared all four procedures within a single context. In addition to producing different performance standards, the discrepancies in most instances were substantial.

The internal consistency of the ratings within a given procedure was extremely stable. Coefficients of .98 and .99 were obtained for all procedures. The validity information, however, did show differences across the procedures. Using item difficulties based on student performance as an external criteria, the correlations with item ratings from each procedure produced moderate coefficients in the .40 to .70 range. The intercorrelations among the Angoff, Ebel and Nedelsky item indices indicated moderate to high coefficients between Angoff and Ebel ratings with low coefficients between Nedelsky and Angoff and Nedelsky and Ebel. Each procedure would appear to be using perceived item difficulty as a basis for judgments, but the specific directions for a procedure tend to alter these perceptions and creates variability in item ratings unique to a procedure.

For any pair of methods, the Ebel and Angoff procedures appear to be most similar in both the pattern of ratings across the same items and in the resultant performance score standard which is produced. Of these two, the Angoff composite ratings exhibited substantially more variability across raters than did the Ebel procedure. The consequence of this variability is that the standard error of the Angoff performance standard would be considerably greater than that of the Ebel procedure. Across groups of similar judges, the Ebel procedure would produce the more stable performance standard.

It is difficult to interpret the error classification rates in Table 7. Comparative interpretation is relative to the validity assumed for the independent teacher judgments of student competency. The evidence that exists suggests that teacher classification is not very highly related to actual student test performance. Correlations between these two indices for Grades 2, 4, 6 and 8 are .51, .62, .60 and .52 in reading, respectively, and .41, .57, .65 and .67 in mathematics, respectively. Given this relationship, it is difficult to justify a decision which sets teacher classification as the ultimate criterion.

The ideal validity study on standard setting procedures would result in the same decisions across all procedures used. However, the data in the present study accentuates the fact that the procedures studied are different rather than similar and produce different results. The conclusion that one is superior to another cannot be drawn without a consensus external criterion. Rather, the information provided on each procedure should offer a basis for a more valid selection of a method if standard setting is desired.

The lack of similarity among procedures supports the position that level of competency needs to be viewed as a continuous variable. If a testing program is oriented toward the purpose of providing information and feedback about a group rather than certification, it would seem desirable to provide performance information for a variety of cut-points, e.g., 90 percent, 80 percent, 70 percent, etc. of the items correct. Those individuals responsible for making policy decisions based on the data may then impose their own internal standards when recommending decisions. The use of a single method to set a performance standard is arbitrary and the existing literature and present data would not support the superiority of any one of the four methods investigated.

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. Thorndik (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1971.

BERK, R.A. A consumers' guide to criterion-referenced test reliability. Journal of Educational Measurement, 1980, 17, 323-349.

BRENNAN, R.L. Applications of generalizability theory. In R.A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

BRENNAN, R.L. & KANE, M.T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289(a).

BRENNAN, R.L. & LOCKWOOD, R.E. A comparison of two cutting score procedures using generalizability theory. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979.

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

FISHER, R.A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7, 179-188.

GLASER, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

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.

HAMBLETON, R.K., POWELL, S., & EIGNOR, D.R. Issues and methods for standard-setting. Amherst, MA: School of Education, University of Massachusetts, 1979.

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

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

References (contd)

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.

POPHAM, W.J. & HUSEK, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

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

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.

Table 1

Number of Students Rated by Teachers
on Competency in Reading and Mathematics
by Grade Level

Teacher Ratings/ Classification		Grades 2	4	6	8
		Reading			
Minimally Competent		1290	1335	1353	1982
Not Minimally Competent		38	93	101	86
	TOTAL	1328	1428	1454	2068
		Mathematics			
Minimally Competent		1299	1340	1307	1923
Not Minimally Competent		18	103	117	190
	TOTAL	1317	1443	1424	2113

Table 2

Distributional Characteristics of the Judges' Ratings For the Angoff, Ebel and Nedelsky Procedures

Angoff (N=312)

	N	$\overline{X}$	Mdn	Q_1	Q_3	S	SK	Ku
R2	37	36.4	37.4	33.9	39.9	4.5	- .9	- .2
R4	32	42.4	45.1	34.4	52.0	11.2	- .5	-1.3
R6	28	43.3	42.8	38.6	57.5	7.8	- .8	.3
R8	35	42.3	44.3	35.6	50.6	9.8	- .7	.1
R11	30	41.5	43.0	37.1	47.0	7.8	-1.1	1.2
M2	32	39.1	39.8	36.8	41.8	3.8	-1.3	2.3
M4	26	45.6	48.6	43.2	51.2	8.7	-1.3	.6
M6	33	42.9	43.6	37.8	57.5	7.5	- .5	- .1
M8	28	38.2	39.3	32.3	45.5	10.1	- .3	- .7
M11	31	36.9	39.3	30.0	44.8	10.3	- .6	- .6

Ebel (N=337)

	N	$\overline{X}$	Mdn	Q_1	Q_3	S	SK	Ku
R2	36	37.3	37.7	36.8	38.9	2.4	- .7	.1
R4	25	42.2	42.0	39.6	44.8	3.9	.2	.1
R6	31	46.3	46.8	44.1	48.2	3.1	- .2	- .7
R8	38	47.3	47.4	45.1	49.5	2.7	- .01	- .8
R11	37	45.0	45.1	43.4	47.1	3.2	- .3	- .8
M2	37	37.4	37.2	36.2	41.3	2.4	-1.0	1.0
M4	33	46.3	46.2	44.4	48.3	3.1	- .3	.1
M6	31	46.8	45.6	44.4	49.3	3.3	.6	-1.1
M8	29	44.9	45.0	43.6	46.7	3.2	- .7	.9
M11	41	43.4	43.1	41.5	44.9	2.4	.6	.8

Nedelsky (N=277)

	N	$\overline{X}$	Mdn	Q_1	Q_3	S	SK	Ku
R2	32	19.1	19.4	16.8	21.3	2.7	- .5	-1.0
R4	23	25.2	24.8	22.8	28.1	3.3	- .2	- .7
R6	24	24.2	23.8	21.3	27.7	3.7	.5	-1.0
R8	30	25.0	25.2	23.2	27.0	2.4	.1	-1.0
R11	29	23.3	23.3	21.5	24.8	2.1	- .3	- .7
M2	32	18.0	18.0	16.2	20.3	2.9	- .1	-1.0
M4	25	25.1	25.3	22.5	27.5	3.6	- .2	-1.1
M6	25	26.5	27.0	25.5	27.7	2.6	- .8	- .1
M8	28	25.0	25.0	23.2	27.0	2.9	- .2	- .6
M11	29	20.5	21.1	17.7	22.9	2.9	- .3	-1.4

Table 3

Internal Consistency Reliability Coefficients
For Judges' Responses Within the Angoff, Ebel and
Nedelsky Procedures

	Angoff	Nedelsky		Ebel			
		Alternative	Composite	Difficulty	Relevance	Cell%	Composite
R2	.98	.97	.97	.96	.95	.89	.98
R4	.99	.93	.97	.94	.97	.90	.98
R6	.99	.98	.99	.93	.94	.93	.98
R8	.99	.97	.98	.90	.92	.93	.98
R11	.97	.98	.98	.94	.99	.94	.99
M2	.96	.96	.98	.93	.91	.91	.98
M4	.98	.98	.99	.90	.93	.91	.97
M6	.98	.96	.98	.93	.99	.95	.99
M8	.99	.98	.99	.96	.95	.91	.97
M11	.99	.99	.99	.92	.95	.95	.98

Table 4

Correlations Among Standard Setting Method Item Ratings and Test Item Difficulties

	METHOD				
Test (# items)	Angoff	Nedelsky	Ebel		
			Difficulty	Relevance	Composite
Reading 2(45)	.66	.52	.63	.25	.47
Reading 4(60)	.66	.56	.73	.25	.71
Reading 6(60)	.49	.47	.51	.56	.28
Reading 8(60)	.38	.52	.15	.17	.40
Reading 11(57)	.46	.31	.45	.19	.42
Math 2(45)	.71	.50	.59	.43	.55
Math 4(60)	.74	.52	.72	.37	.56
Math 6(60)	.81	.48	.73	.56	.65
Math 8(60)	.41	.28	.19	.24	.62
Math 11(57)	.54	.46	.48	.29	.55

Table 5

Correlations Among Angoff, Ebel and Nedelsky Item Ratings

Test	Angoff with				Nedelsky with		
	Nedelsky	Ebel(Diff)	Ebel(Rel)	Ebel(Comp)	Ebel(Diff)	Ebel(Rel)	Ebel(Comp)
R 2	.718	.937	.727	.865	.818	.565	.802
R 4	.778	.924	.473	.894	.715	.468	.706
R 6	.545	.925	.635	.635	.623	.327	.403
R 8	.542	.196	.647	.850	.152	.105	.388
R11	.559	.886	.398	.734	.660	.034	.370
M 2	.483	.942	.742	.902	.437	.312	.350
M 4	.626	.897	.701	.827	.740	.252	.503
M 6	.510	.888	.687	.866	.401	.202	.359
M 8	.123	.914	.821	.561	.061	.021	.237
M11	.454	.850	.176	.740	.477	.035	.509

Table 6

Indices of Dependability for Minimum Passing Scores Suggested by Each Procedure

Criterion Score	Reading					Mathematics				
	2	4	6	8	11	2	4	6	8	11
18	.992	.992	.989	.993	.995	.997C	.993	.992	.990	.984
19	.991	.991	.988	.992	.994	.997	.993	.991	.990	.983
20	.990	.991	.987	.992	.994	.996	.993	.990	.989	.982
21	.989	.990	.986	.991	.994	.996N	.992	.990	.988	.980
22	.988N	.989	.985	.990	.993	.995	.992	.989	.987	.978
23	.987	.988	.984	.990	.993	.995	.991	.988	.986	.977
24	.986	.988	.983	.989	.992	.994	.990	.987	.985	.975N
25	.984	.987	.982	.988	.992	.994	.990	.986	.984	.972
26	.982	.986	.980	.987	.991N	.993	.989	.985	.983	.970
27	.979C	.985	.978	.986	.990	.992	.988	.984	.982	.967
28	.976	.983	.976N	.985N	.989	.991	.987	.983	.980N	.964
29	.973	.982N	.974	.984	.988	.990	.986N	.981	.978	.960
30	.969	.980	.972	.982	.987	.989	.985	.980N	.977C	.956
31	.963	.979	.968	.981	.986	.987	.984	.978	.979	.952
32	.957	.977	.966	.979	.984	.984	.982	.976	.972	.947
33	.949	.974	.962	.976	.983	.981	.980	.974	.970	.942
34	.940	.972	.958	.974	.981	.977	.979	.971	.967	.937
35	.929	.969	.953	.971	.979	.972	.976	.968	.964	.931
36	.917	.965	.948C	.968	.976	.965	.974	.965	.960	.926
37	.905A	.962	.942	.964	.973	.956	.971	.961	.956	.921
38	.894E	.957	.935	.959	.969	.943E	.968	.957C	.952	.917
39	.887	.952	.928	.954C	.964	.924	.965	.952	.948A	.914
40	.886	.947	.920	.948	.958	.900A	.961	.948	.944	.912A
41	.891	.941	.913	.940	.951	.872	.957C	.942	.940	.912
42	.901	.935C	.905	.932	.942A	.850	.952	.937	.936	.914
43	.913	.929A, E	.898	.923A	.931	.846	.947	.932A	.933	.916
44	.926	.922	.892A	.912	.918	.864	.942	.927	.930	.920E
45	.937	.916	.889	.901	.902E	.891	.937	.922	.928E	.925
46		.911	.888	.890	.882		.932A	.919	.928	.930
47		.908	.889E	.880	.861		.928E	.917E	.928	.936
48		.907	.893	.872E	.840		.925	.917	.930	.941
49		.907	.899	.867	.823		.924	.918	.933	.946
50		.910	.906	.867	.815		.923	.921	.936	.951

A = Angoff Standard
C = Contrasting Groups Standard

E = Ebel Standard
N = Nedelsky Standard

Table 7

Classification Agreement Based on Performance Standards for Each Procedure Using Teacher Judgments as the Criteria

Teacher Classification	METHOD Nedelsky N-M	 M	 Contrast. Groups N-M	 M	 Angoff N-M	 M	 Ebel N-M	 M
Reading 2:								
Non-Masters	4	34	14	24	29	9	29	9
Masters	5	1285	15	1275	184	1106	320	1060
Standard	22		27		37		38	
Reading 4:								
Non-Masters	32	61	68	24	74	19	74	19
Masters	27	1308	134	1201	152	1183	152	1183
Standard	29		42		43		43	
Reading 6:								
Non-Masters	34	67	66	35	84	17	93	8
Masters	32	1321	134	1219	352	1001	497	856
Standard	28		36		44		47	
Reading 8:								
Non-Masters	8	78	25	61	34	52	48	38
Masters	10	1972	75	1907	149	1833	408	1574
Standard	28		39		43		48	
Mathematics 2:								
Non-Masters	0	18	0	18	10	8	8	10
Masters	1	1298	0	1299	94	1205	42	1257
Standard	21		0		40		38	
Mathematics 4:								
Non-Masters	0	103	69	34	83	20	87	16
Masters	24	1316	149	1191	245	1095	267	1073
Standard	29		42		46		47	
Mathematics 6:								
Non-Masters	51	66	92	25	102	15	109	8
Masters	46	1261	146	1161	267	1040	426	881
Standard	30		38		43		47	
Mathematics 8:								
Non-Masters	57	133	70	120	122	68	137	53
Masters	50	1873	64	1859	300	1623	563	1360
Standard	28		30		39		45	

Table 8

Descriptive Statistics Found for the Kansas Competency Tests

Area	Grade	Items	$\overline{X}$	Mdn.	S	$\overline{P}$	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
Reading	11	57	50.1	51.5	5.6	.88	30,881
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
Mathematics	11	57	40.6	42.3	10.6	.71	30,752