DOCUMENT RESUME

ED 414 292		TM 027 721
AUTHOR	Sireci, Stephen G.; Robin, Frederic; Patelis, Thanos	
TITLE	Using Cluster Analysis To Facilitate the Standard Setting Process.	
PUB DATE	1997-03-27	
NOTE	39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 24-27, 1997).	
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)	
EDRS PRICE	MF01/PC02 Plus Postage.	
DESCRIPTORS	*Cluster Analysis; *Cutting Scores; *Educational Testing; Evaluation Methods; Grade 7; *Junior High School Students; Junior High Schools; Mathematics Tests; *Pass Fail Grading; Standards	
IDENTIFIERS	*Standard Setting	

ABSTRACT

The most popular methods for setting passing scores and other standards on educational tests rely heavily on subjective judgment. This paper presents and evaluates a new procedure for setting and evaluating standards on tests based on cluster analysis of test data. The clustering procedure was applied to a statewide mathematics proficiency test administered to 818 seventh-grade students in a small urban/suburban school district. Content area subscores were derived from the test specifications to serve as clustering variables. Subsequent course grades in mathematics were used to validate the cluster solutions and the stability of the solutions were evaluated using two random samples. The three-cluster (K-means) solution provided relatively homogeneous groupings of students that were consistent across the two samples and were congruent with school mathematics grades. Standards for "intervention," "proficient," and "excellent" levels of student performance were derived from these results. These standards were similar to those established by the local school district. The results suggest that cluster analytic techniques may be useful for helping set standards on educational tests, as well as for evaluating standards set by other methods. Suggestions for future research are provided. (Contains 2 figures, 7 tables, and 23 references.) (Author/SLD)

Using Cluster Analysis to Facilitate the Standard Setting Process

Stephen G. Sireci and Frédéric Robin
University of Massachusetts at Amherst

Thanos Patelis
Stamford Public Schools

Paper presented at the annual meeting of the National Council on Measurement in Education, March 27, 1997, Chicago, Illinois.

Using Cluster Analysis to Facilitate the Standard Setting Process

Abstract

The most popular methods for setting passing scores and other standards on educational tests rely heavily on subjective judgment. This paper presents and evaluates a new procedure for setting and evaluating standards on tests based on cluster analysis of test data. The clustering procedure was applied to a statewide mathematics proficiency test administered to seventh grade students in a small urban/suburban school district. Content area subscores were derived from the test specifications to serve as clustering variables. Subsequent course grades in mathematics were used to validate the cluster solutions and the stability of the solutions was evaluated using two random samples. The three-cluster (K-means) solution provided relatively homogeneous groupings of students that were consistent across the two samples and were congruent with school mathematics grades. Standards for "intervention," "proficient" and "excellence" levels of student performance were derived from these results. These standards were similar to those established by the local school district. The results suggest that cluster analytic techniques may be useful for helping set standards on educational tests, as well as for evaluating standards set by other methods. Suggestions for future research are provided.

Introduction

Standard setting, also known as setting "cutscores" on tests, is an important and enigmatic problem in psychometrics. The past several decades have witnessed several technological developments in testing, such as new theories and methods for developing, scaling, scoring, and administering educational and psychological tests. Unfortunately, there have been few widely embraced innovations in standard setting. The lack of acceptable methods for setting standards is troubling because many tests involve standards, and important consequences are tied to them. For example, in employment or licensure testing, scoring at or above a standard may make the difference in earning a job, promotion, or license. When cutscores are used on tests, standard setting becomes a critical validity issue. It directly impacts the validity of inferences derived from test scores, and is related to all issues of test fairness and utility.

This paper presents a new methodology, based on cluster analysis, that can be used to help set standards on educational tests. Like all standard setting procedures, the standards suggested by the cluster analytic method are not absolute. Therefore, this method does not "solve" the standard setting problem, but it does help inform the process.

Setting Discrete Standards on Continuous Score Scales

Part of the problem in setting standards on educational and psychological tests is that the test development and scaling processes typically order test takers along a *continuous* scale. Test developers have become quite proficient at constructing scales with a pre-specified mean and standard deviation, and at estimating the distances among test takers along this scale. The most common scaling practice is to use the standard deviation of observed scores as the unit of measurement for expressing the distance of test takers from the mean of the score scale.

Assuming a normal distribution of the characteristic measured, test scores may be scaled by computing a standard normal deviate of the form

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

where z is the standard normal deviate corresponding to score x, and μ and σ are the mean and standard deviation of a norm group, respectively. Typically, these standard normal deviates are transformed to a scale with a pre-specified mean and standard deviation. These processes result in a continuous score scale. However, the standard setting problem is *discrete*. When cutscores are used to classify test takers into one or more groups (e.g., distinguish between "passers" and "failers"), score differences among examinees within each group are typically inconsequential. When standards are set on tests, the fundamental scaling problem is not how to best order examinees along a continuous scale; it is how to best partition test takers into the desired number of (discrete) groups motivated by the purpose of testing.

Many test specialists have pointed out the problem of forcing discrete decisions on a continuous score scale (e.g., Dwyer, 1996). Unfortunately, the use of a discrete scaling process to classify test takers has not been widely investigated. One method for ordering examinees along a discrete scale is to use cluster analysis. Cluster analysis could be used to group test takers into homogeneous clusters with respect to the proficiency measured. Each cluster would comprise examinees highly similar in proficiency. These clusters could then be ordered in a manner congruent with the a priori groupings defined by the standard setting problem. The logic underlying this method is explained further in the next section. Subsequently, the results of an application of this method are presented and critiqued.

Scaling Test Takers Using Cluster Analysis

Cluster analysis could be used in several ways to classify test takers into meaningful and qualitatively different proficiency groupings. Using a partitioning method such as K-means clustering, distances among homogenous groups of examinees (clusters) could be computed by determining the distance of each test taker to the center of each cluster, The center of each cluster is represented by a vector of means (called the cluster centroid) corresponding to the variables used to cluster the test takers. Each test taker is assigned to a cluster by computing the distance between the test taker and each cluster, and assigning her/him to the "closest" cluster. This type of scaling has two obvious differences from traditional psychometric scaling. First, the distances among test takers is not determined from a single mean, but rather from a vector of means. Second, rather than placing test takers on a continuous scale, test takers are placed into one of a discrete number of clusters. The standard setting problem then becomes identifying which clusters correspond to the proficiency groupings invoked by the standard setting and test development processes.

A meaningful cluster analysis solution could inform the standard setting process in several ways. For example, the investigator could examine the cluster solution in search of a cluster of "borderline" examinees. If such a cluster is found, the scores of these borderline test takers could be used to derive a cutscore (e.g., the median test score of these borderline examinees becomes the cutscore). Inspecting such clusters may also prove helpful for envisioning or determining borderline examinees as they are conceptualized in Angoff-type standard setting studies. Alternatively, the cluster solution could be inspected to determine whether specific clusters of examinees emerge such as "passers," "failers," "masters," etcetera.

The logic underlying the use of cluster analysis for setting standards is clear and intuitively appealing. However, the procedure has two conspicuous limitations. First, because it focuses on analysis of test response data, no standards can be set higher or lower than the test takers actually performed. The procedure will suggest standards based on what test takers *can do*, rather than according to what they *should do*. Although this limitation is serious theoretically, it is unlikely that a test would be constructed so far above or below examinee performance levels that *no* test takers exhibit expected standards of performance. The second obvious limitation stems from the fact that clustering procedures will cluster the data regardless of whether truly different groups of examinees are present. For this reason, an external criterion (i.e., a variable not used to cluster objects, but related to true cluster structure) is always needed to validate the cluster solution, and help justify that the resulting clusters are qualitatively different from one another.

It should also be noted that unlike traditional applications of cluster analysis, when cluster analysis is used to inform the standard setting process, the goal is not to uncover the "true" cluster structure of the data. Rather, the goal is to identify the optimal partitioning of the examinee population that best corresponds to desired a priori proficiency groupings.

Method

Instrument

The Connecticut Mastery Tests

The test data analyzed come from a non-mandated component of the Connecticut Mastery Testing Program, which is a statewide basic skills testing program in reading, writing, and mathematics. The primary purposes of the Connecticut Mastery Tests (CMT) are to identify students in need of remedial instruction and to monitor students' attainment of expected standards of achievement. All public school students in Connecticut entering fourth, sixth, and eighth grades are required to take these tests by the Connecticut State Department of Education (CSDE). The CMT is designed to assess how well each student is performing with respect to skills identified as important for them to have mastered. Three standards of performance are established in grades four, six, and eight to identify "intervention," "proficient," and "excellence" levels of performance. These standards were established by the CSDE using a modified Angoff[1] standard setting procedure (CSDE, 1994). Voluntary editions of the CMT were also developed by the CSDE for students in third, fifth, and seventh grades. However, State standards are not available at these grade levels. Thus, if they choose to use these non-mandated tests, local school districts are encouraged to develop their own standards for intervention, proficient, and excellence levels of performance.

This study analyzed test data from the grade seven mathematics section of the CMT administered to students in a local urban/suburban school district. This mathematics test comprised 134 items; 110 of the items were multiple-choice, 8 required examinees to "bubble in" a numerical answer, and 16 were open-ended requiring completion of a problem. Nine of the open-ended problems were scored on a three-point scale; all other test items were scored dichotomously (right/wrong). Thus, the total scores on the test could range from zero to 143. The test measured 34 mathematics objectives comprising four global content areas: concepts (24 items), computation and estimation (36 items), problem solving/applications (54 items), and

[1]The modified Angoff standard setting procedure involves the use of subject matter experts to inspect all test items and provide judgments regarding the probability of success (or expected score) of "borderline" test takers on each item. Standards are derived by averaging these ratings over judges. See Cizek (1996) or Livingston and Zieky (1982) for more complete details.

measurement/geometry (20 items).

To establish the intervention, proficient, and excellence standards on this test, the local board of education used interpolated percent correct scores (percentage of items answered correctly) that best matched the CSDE cutscores on the mandated sixth and eighth grade mathematics examinations (established using the modified Angoff procedure). For example, if the percent correct score for the excellence standard was 88% for the eighth grade test and 90% for the sixth grade test, the locally-established standard for excellence was the score that represented a percent correct score of 89%. Using this process, the cutscores of 71 and 112 were used to distinguish between intervention/proficient and proficient/excellence levels of performance, respectively. Using this process, the percentages of students classified in each of the three performance levels was similar across the three grade levels.

Subjects

The data for all seventh grade students tested in the school district were analyzed (n=818). Eight students who earned a raw score below 27 (which is the score expected by guessing alone) were eliminated from the analysis because these scores may signify non-serious test taking behavior and could adversely affect the cluster solutions. The final sample comprised 810 examinees. The internal consistency reliability of the scores for these students was high (α=.97), which was expected given the large number of test items. The mean raw test score was 93.3 and the standard deviation was 29.3. Using the standards described above, 27.2% of students in the data set scored in the intervention category, 40.2% scored in the proficient category, and 32.6% scored in the excellence category.

To evaluate the stability of cluster solutions across different samples of examinees, two random samples (without replacement) were created from the original data file. Each sample contained 405 examinees, and are referred to as sample A and B, respectively.

Procedure

Cluster variables

There are at least three options for selecting the variables to be used to cluster test takers: 1) use all individual items comprising the test, 2) use orthogonal factor scores obtained from item-level factor analysis, or 3) use subscores derived from items comprising the major content areas of the test. Given the large number of items comprising the test, the high inter-correlation among the content areas, and suggestions based on previous research (Milligan, 1995; Sireci & Robin, 1996; Sneath, 1980), the third method based on content area subscores was used. Subscores for each of the four content areas measured by the test (concepts, computation and estimation, problem solving/applications, and measurement/geometry) were used as the input variables for all cluster analyses. These subscores were computed for all students by summing their item scores within each content area. This strategy modeled the test content specifications, which reflected the content areas deemed important by the curriculum specialists who designed the test. The four subscores were highly correlated ranging from .75 (correlation between concepts and measurement/geometry) to .88 (correlation between concepts and problem solving). The content area subscores were standardized (μ=0, σ=1) prior to clustering to account for differences in the raw score scales due to the different number of items in each content area (20 to 54 items).

Data preliminaries

The distribution of test scores was inspected visually to determine if modes or gaps existed in the data that correspond to proficiency groupings of examinees. A histogram of test scores is presented in Figure 1. As is typical of the standard setting problem, no clear gaps appear in the data. Scatterplots indicating the relationship among the subscores were also evaluated to determine the most appropriate type of clustering method to use. No well-separated clusters were observed, although there were some regions of the multidimensional space that contained relatively tight groupings of examinees.

[Insert Figure 1 About Here]

Clustering procedures

Both hierarchial and K-means cluster analysis procedures were used in this study. The hierarchical procedures were conducted as a preliminary step in determining the number of clusters to fit to the data in the subsequent K-means analyses. In all cluster analyses, the distances among students, or among students and clusters were computed using the Euclidean distance formula:

$$d_{ij}=\sqrt{\sum_{a=1}^{r}(x_{ia}-x_{ja})^2} \qquad (2)$$

where d_{ij} is the distance between two entities i and j (such as between examinees i and j, or between examinee i and cluster j), x_{ia} is the score of examinee i on measure a, and r is the total number of measures used to form the clusters.

Because K-means cluster analysis requires specification of the number of clusters in advance, hierarchical cluster analyses (HCA) were initially performed on each sample to help determine the approximate number of clusters in the data. Euclidean distances were computed between cases (based on the four content area subscores), and Ward's (1963) method was used to cluster examinees. Ward's method was chosen because it tends to produce relatively dense clusters, which were expected due to the relatively tight groupings of examinees observed in the multivariate content area scatterplots. HCA begins by treating each student as a separate cluster and sequentially merges students into clusters until all students comprise a single cluster. The task for the investigator is to determine the stage at which clusters that are different are merged. The C-index (Dalrymple-Alford, 1970; Hubert & Levin, 1976) was used to help determine the appropriate number of clusters in the HCA and K-means solutions. This descriptive index reflects the internal cohesion of the cluster solution and has performed well under simulation conditions when the true number of clusters was known (Milligan, 1980, 1981; Milligan & Cooper, 1985). The equation for the c-index is

$$C = \frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)} \tag{3}$$

where d_w is the sum of within-cluster distances across all clusters comprising a given solution (i.e., sum of all distances of each student from the cluster centroid to which s/he belongs), max(d_w) is the sum of distances for the cluster whose within-cluster distances are largest (i.e., the most heterogeneous cluster of a solution), and min(d_w) is the sum of distances for the cluster whose within-cluster distances are the smallest (i.e., the most homogeneous cluster of a solution). The minimum value of the C-index calculated for competitive cluster solutions indicates the "best" solution in terms of minimizing within-cluster distances and maximizing between cluster distances. However, this index is a descriptive statistic, and does not always reveal the true cluster structure.

A disadvantage of HCA is that early inefficient cluster assignments cannot be corrected later in the algorithm. For this reason, the HCA analyses were considered preliminary and a K-means procedure was used to partition the students. The K-means algorithm is iterative and so students can be shifted from one cluster to another until a "min/max" convergence criterion is reached (minimizing within cluster distances and maximizing between cluster distances). K-means cluster analysis requires a priori specification of the number of clusters fit to the data. The Q-cluster solution (where Q represents the number of clusters) that resulted in the minimum value of the C-index for the HCA provided an estimate of the number of homogeneous clusters in the data. Subsequently, Q-2 to Q+2 K-means cluster solutions were obtained.

The "K" in K-means clustering corresponds to the number of variables used to cluster the test takers, so in this study K=4. The centroid of each cluster is defined by a vector of the four means, and distances of the students to the cluster centroid were determined using equation 2. The hierarchical and K-means cluster analyses were conducted using SPSS for Windows, version 6.0 (SPSS, 1993). The K-means algorithm within SPSS uses Anderberg's (1973) nearest centroid sorting method. The convergence criterion used was a maximum distance change of less than .001 by any cluster center between two sequential iterations.

Validating the cluster solutions

The final grades obtained by the students in their mathematics courses were used as external criteria to validate the cluster solutions. Complete grade data were available for 774 of the 810 students (96%). The final grades were reported on a five-point scale ranging from "F" to "A;" however, there were three educational tracks within three of the schools. To account for the different educational tracks, two points were added to the grades of the students in the accelerated ("high") track and one point was added to the grades of the "typical" track students. Thus, the recoded grades ranged from one to seven. The fourth school in the sample was a magnet school that did not use separate educational tracks. The grades for these students were adjusted by one point to fall in the "typical" educational track range. Although this recoding scheme is not perfect, it provided improved information regarding the mathematics achievement of the students in comparison to the unadjusted grades. An illustration of how the grades were recoded is presented in Table 1. The correlation between total scores on the test and the recoded mathematics grades was .72. Although recoded or unadjusted school grades are not a perfect achievement criterion, they do represent an independent mechanism for describing differences in mathematic achievement among the students that is useful for evaluating the cluster solutions.

[Insert Table 1 Here]

To evaluate the cluster solutions, one-way ANOVAs were conducted on the K-means solutions using cluster membership as the independent variable and school mathematics grade as the dependent variable. This analysis represented an external evaluation of the solutions because grades were not used as a variable to cluster examinees. In addition, cross-tabulations were computed between cluster membership from the cluster solution and final grade. The relationship between cluster solution membership and recoded mathematics grades was used to help determine the most appropriate cluster solution. Cluster solutions that partitioned the students in a manner congruent with their final grades were considered to be valid. Contrariwise, cluster solutions that partitioned students with similar final grades were considered to reflect an "over-clustered" solution. A further criterion for evaluating the solutions was stability across the two samples in terms of cluster profiles.

Setting the standards using cluster analysis

Descriptive test score statistics were computed separately for each cluster within a solution. The median, minimum, and maximum test scores within each cluster were used along with the cluster centroids to determine whether an observed cluster of students could be classified as intervention, proficient, excellence, or borderline. It was hypothesized that establishing the passing standard would be dependent upon the type of cluster solution observed. If the solution clearly separated two groups, such as intervention students and proficient students, the passing score would be set somewhere between the maximum and minimum test scores of these two groups, respectively. If a cluster of "borderline" students was observed, it was hypothesized that the median value of the total test score for these students could be used as the passing standard. To evaluate the appropriateness of where the standards were placed, the correlation between classifications based on the cluster analysis and average course grade were computed. These correlations served as an index of the congruence between classification decisions made on the basis of the cluster analyses and those made by the teachers (i.e., teachers' grades).

Results

Hierarchical Cluster Analyses

The C-indexes computed for the two- through six-cluster HCA solutions for both samples of students are presented in Table 2. The results were equivocal across the two samples. The two- and four-cluster solutions exhibited the smallest C-indexes for sample A, and the three-cluster solution provided the smallest C-index for sample B. Based on these results, subsequent 2- through 6-cluster K-means solutions were computed for both samples.

[Insert Table 2 About Here]

K-means Cluster Analyses

C-indexes were also calculated to compare the K-means cluster solutions. These results are presented in Table 3. The C-index was not consistent across the two samples. It was lowest for the 6-cluster K-means solution for sample A, and the 4-cluster K-means solution for sample B. Thus, the criterion of internal cohesiveness was not particularly helpful for determining the best cluster solution.

[Insert Table 3 About Here]

Descriptive statistics for the K-means cluster solutions are presented in Table 4. The second column in Table 4 lists the clusters in each solution and indicates whether it is from sample A or B. The third column gives the proportion of the 405 students in the cluster. The next three columns give the median, minimum, and maximum test scores for the students in each cluster. The last four columns report the cluster centroids, which are the means on each of the standardized content area subscores for all students within the cluster. In evaluating the stability of the cluster solutions across the two samples (in terms of cluster centroids and descriptive test score statistics), consistency is most evident for the two- and three-cluster solutions. The similarity across samples tended to decrease as the number of clusters in the solution increased. The median test scores for the different groups within a cluster solution are very similar across samples A and B for the two- and three-cluster solutions, as are the minimum and maximum scores.

[Insert Table 4 About Here]

The results of the one-way ANOVAs, which tested for significant differences in school math grades among all clusters within a solution, are presented in Table 5. The last column in Table 5 indicates whether all pairwise comparisons between clusters within a solution were statistically significant using the Scheffé post-hoc comparison procedure, with a within-cluster (familywise) alpha equal to .05. The conservative Scheffé procedure was chosen due to the relatively large sample sizes. Although the power of detecting a statistically significant difference decreases as the number of clusters within a solution increased, the sample sizes for clusters within the higher-cluster solutions were considered sufficient for detecting a statistically significant difference. [The smallest cluster (sample) sizes for the four-, five-, and six-cluster post hoc analyses were 79, 64, and 38, respectively.] For both samples A and B, the clusters arising from the two-, three-, and four-cluster solutions were statistically significantly different from one another in terms of recoded math grades. All pairwise comparisons were not statistically significant for the five- and six-cluster solutions. These results suggest that additional clusterings beyond the four-cluster solution are segregating students that may not be very different from one another in terms of school math grades.

[Insert Table 5 About Here]

Given the variability of the C-indexes across samples, evaluation of the cluster stability across samples, and the ANOVA results, the two-, three-, and four-cluster solutions appear appropriate for these data. Comparisons of the maximum and minimum scores between the lower and higher proficiency groupings indicated that the three- and four-cluster solutions would provide intervention/proficient classifications in a manner more consistent with cluster membership than would the two-cluster solution. In comparing the three- and four-cluster solutions, it appears that four-cluster solution adds a "borderline" cluster between what may be considered the intervention cluster and the proficient cluster. Although either the three- or four-cluster solution could be used to determine the cutscores, the three-cluster solution was chosen because the number of clusters in the solution matched the number of proficiency groupings for which the test was designed to identify. Thus, focusing on the overlap among these three clusters paralleled the potential overlap among the proficiency groupings used by the school district. In addition, the three-cluster solution exhibited greater similarity across samples A and B, in comparison to the four-cluster solution.

To evaluate how well the three-cluster solution grouped students according to their rescored mathematics grades, cluster membership was cross-tabulated with rescored grade, and the correlation between the three ordered proficiency clusters and rescored grade was calculated. Table 6 presents the cross-tabulation separately for sample A and B. Across both samples, no students with rescored grades of "7" (the highest possible grade) were grouped into the "intervention" cluster, and no students with grades less than "3," were grouped into the "excellence" cluster. The Spearman rank-order correlation between the ordered clusters and rescored grades was .69 for both Samples A and B. Given the different educational tracks noted across the schools, it appears that the three-cluster solution grouped the students well in terms of their in-school mathematics achievement.

[Insert Table 6 About Here]

Setting the cutscores

Using the three-cluster solution, the minimum score for the "middle" proficiency cluster (cluster 2) was 73 for sample A and 72 for sample B, while the maximum values for the lowest proficiency cluster (cluster 1) were 75 and 78, respectively. Thus, the score interval 72-to-78 appears to be the best place to establish the intervention/proficient cutscore. The maximum score values for the middle proficiency cluster were 112 and 111 for samples A and B, respectively. Looking at the highest proficiency cluster (cluster 3), the minimum scores achieved by these students were 107 and 104 for the two samples. Thus, it appears the best location for the proficient/excellence cutscore is somewhere within the 104-to-112 score interval.

To derive sensible cutscores within these two intervals, samples A and B were combined and the median total test score for examinees *within these two intervals* were calculated. The median test scores of students scoring in the lower interval was 75. Thus, a test score of 75 was used as the intervention/proficient cutscore. The median test score of students scoring in the upper interval was 107, which was used as the proficient/excellence cutscore. These two cutscores are close to those established by the local school district for this test (71 and 112, respectively). In comparing the two sets of cutscores, the cutscores derived via cluster analysis classified 3.8% more students as intervention, 9% fewer students as proficient, and 5.2% more students as excellence than did the cutscores established by the local board of education. Thus, using the cluster-analytic derived cutscores, fewer students meet the proficient standard, and more students meet the excellence standard.

To evaluate these standards empirically, the standards derived from the cluster-analytic solution were compared to those established at the local level. First, the students were categorized into one of the three proficiency groups according to the cutscores derived by the local school district. The Spearman rank-order correlation between these locally-derived groupings and those derived using the clustering procedure was .95. Next, Spearman correlations between the classifications made based on the different standard setting procedures and course grades were computed. The standards derived using cluster analysis correlated .69 with the rescored grades; the correlation observed for the standards established at the local level was .68. In addition, the classifications resulting from these two different methods (i.e., using the two different sets of cutscores) exhibited similar correlations with the total test score (.94 for the cluster-analytic standards and .93 for the locally-established standards). In general, the standards set using the two different methods exhibited similar relationships with the rescored grades and total test scores.

Using the graphing procedures suggested by Friedman and Rubin (1967), the cluster memberships for the sample A students are presented visually in Figure 2. This figure plots the students in the two-dimensional space determined by their coordinates on two discriminant functions. The functions were calculated by performing a discriminant analysis using the content area subscores to predict cluster membership. The Friedman-Rubin procedure circumvents the difficulty of plotting students in the original four-dimensional content area subscore space. As is evident in Figure 3, there is essentially no overlap across clusters. However, students alongside the "borders" of adjacent clusters are very close to one another. Only sample A students are plotted to make the plot easier to decipher. The scatterplot for sample B (not shown) tells essentially the same story. The Friedman-Rubin plot indicates that there is little or no overlap among the clusters with respect to test performance, but that the maximum and minimum students in adjacent clusters are very similar to one another. These students who are close to one another, but belong to different clusters (located primarily above the -2 and +2 points on the x-axis in Figure 2) represent those "borderline" students whose median test scores were used to derive the cutscores.

[Insert Figure 2 About Here]

Discussion

This study represents the second evaluation of the utility of cluster analysis for helping set standards on educational tests. Sireci (1995) found that cluster analysis was useful for evaluating the passing standard established on a high-stakes writing skills test used to award high school equivalency degrees to adults taking the Tests of General Educational Development. This study extended the previous work by evaluating a test involving two cutscores, and evaluated the cluster solutions using a more defendable external criterion (actual course grades rather than self-reported grades). Taken together, the results of these two studies indicate that cluster analysis appears useful for both evaluating currently established standards and for helping determine where standards should be set on educational tests. However, research in this area is just beginning and it is clear further study is warranted.

The results of this study suggest that cluster analysis may be a useful tool for helping decide where cutscores should be set. However, like all standard setting methodologies, cluster analysis does not provide "absolute" information regarding the "best" cutscore. As the results of this study illustrate, standard setters have at least three difficult tasks when using cluster analysis to help establish cutscores. First, they must identify meaningful cluster solutions; second, they must compare solutions and select the one that appears most appropriate for setting standards; and third, they must determine how to best use the accepted solution for establishing cutscores. In this study, a three-cluster solution was chosen, and the median scores of students from different clusters who were adjacent to one another were used to set the two cutscores. These decisions seemed sensible given the purpose of the testing and the criteria of replicability and external validation. However, other decisions were possible and defensible (e.g., using the median scores of students comprising cluster number 2 from the four-cluster solution as the intervention/excellence cutscore). Although a three-cluster solution was used in this study to partition the students into the three desired groups, the number of clusters in a solution does not necessarily need to equal the number of groups invoked by the standard setting process. For example, Sireci (1995) used a four-cluster solution for the purpose of evaluating pass/fail decisions.

Setting standards by clustering examinees offers some advantages over methods dependent solely on ratings from subject matter experts (SMEs). Two major criticisms of the SME-based procedures are lack of stability of the standards across different groups of SMEs, and questions regarding whether SMEs are actually able to make comprehensive judgments regarding the probable performance of "borderline" examinees (Angoff, 1988; Cizek, 1993; Plake, 1996; USGAO, 1993). Although methods relying on SMEs are currently the most popular standard setting methods, they have not been able to address these criticisms. The cluster analytic procedure does not suffer from these limitations because the stability of the cluster solution can be evaluated directly by replicating the analyses over two or more samples, complex ratings of test items are not required, and subjective opinions regarding hypothetical borderline examinees are not required. Instead, the cluster solutions are useful for actually discovering who the borderline examinees are. For example, the three-cluster solution accepted in this study revealed three distinct groups of students: one group relatively low across all four content areas, one group relatively high across all four content areas, and one group in between these two extremes (see the last four columns of Table 4). Future applications of this procedure could employ SMEs to assist in evaluating the cluster solutions and help determine the clusters that correspond to meaningful proficiency groupings.

Clustering Versus Strict Criterion-referencing

The cluster analytic approach also holds several advantages over approaches that rely entirely on criterion data from a subset of the examinee population (Livingston & Zieky, 1982, p. 51). For example, a strict external criterion approach could be applied to the mathematics test analyzed here by retrospectively using the median test score of students in the remedial track for the intervention/proficient classification, and the median test score for the high track students as the cutscore for the proficient/excellence classification. Applying this method to the present data uses only 31.1% of the original sample (5.7% for the intervention/proficient classification and 25.4% for the proficient/excellence classification).

The strict criterion-referenced method described above was compared with the cluster analytic method. The median test score for the remedial students was 53.5, and the median test score for the high track students was 127. Using these medians as cutscores, the resulting classification decisions exhibited a Spearman correlation of .57 with the rescored grades, which was much lower than those observed for the cluster-analytic and locally-derived methods (.69 and .68, respectively). The classification decisions based on using these medians as cutscores classified only 10.1% of the students as intervention and only 17.4% of the students as excellence. A comparison of the percentage of students that would be placed in each category by the three different standard setting methods is presented in Table 7. This comparison illustrates that the cluster analytic procedure identified more students as intervention and excellence than did the other two procedures. The strict criterion-referenced approach classified the fewest examinees in these two categories. If the CMT is used to refer students for consideration of placement into remedial or advanced instruction, then the cost of referring a student who may not need the instruction is far less than the cost of not referring a student who may benefit from special instruction (assuming students are further evaluated after referral). Thus, the classification decisions based on the cluster analysis seem best, given the primary purpose of the assessment.

[Insert Table 7 About Here]

Aside from the difference in the students who are classified into the three proficiency groupings, there are at least two other factors supporting use of the cluster analysis procedure over an entirely criterion-referenced procedure. First, is the difference in the number of examinees who are used to establish the passing standard. The clustering procedure utilized the entire sample, whereas the strict-criterion method ignored 69% of the examinees. Second, and more importantly, the criterion method places complete faith in the external criterion. This conviction would not be a problem if perfect criterion data were available. Unfortunately, such criteria rarely exist in practice. In fact, if teachers' course grades could be used to make reliable classification decisions, tests like the CMT would not be needed in the first place. In contrast, the cluster analysis procedure models the content specifications of the test, which represent the mathematics content areas deemed important by the curriculum specialists involved in developing the test. The cluster analysis method does not use criterion data in establishing the standard, rather, it uses criterion data to help identify the most appropriate cluster solution. In so doing, the procedure provides validity information for both the test and criterion.

Correlation of Cluster Variables

The results suggest that subscores derived from the test based on global content areas (the same method used by Sireci, 1995) are appropriate and useful for clustering examinees. It does not appear that the high intercorrelation among the subscores led to a misleading cluster solution. However, to evaluate whether uncorrelated cluster variables would provide better results, principal components were extracted from the inter-item correlation matrix and factor scores on these components were computed for the students. The K-means cluster analyses using these uncorrelated factor scores provided results that were very different, and much less useful, than those derived using the content area subscores. The clusters derived using factor scores exhibited substantial overlap across clusters and were not strongly related to meaningful proficiency groupings of students. Therefore, for these data, using uncorrelated subscores did not improve clustering results. However, future applications of cluster analysis should consider factor scores as an option to content area scores for clustering examinees, especially when the assessment may measure several disparate knowledge and skill areas.

Performance of the C-index

The grade data used as external criteria were given more weight in evaluating the cluster solutions than were the C-indexes. The C-index was useful for providing a range of potential cluster solutions. However, its efficiency in determining the true cluster structure of simulated data (e.g., Milligan, 1981; Milligan & Cooper, 1985) was not observed for the present data. The differences between the nature of the current and simulated data probably explains this finding. The current study involved partitioning a large, continuously distributed sample of objects, whereas the simulation studies have used much smaller sample sizes and data sets involving non-overlapping clusters. For these reasons, the external criterion data were deemed more appropriate for selecting the best cluster solution.

Implications for Standard Setting

This study took a different perspective on standard setting from that currently found in the psychometric literature. When standards are set on educational tests, classification decisions are made, regardless of whether true, homogeneous clusters exist in the examinee population. The results of this study reveal that cluster analysis of the examinee population provides new information regarding where the passing score should be set. The cluster analysis approach does not completely remove subjectivity from the standard setting process, but rather provides SMEs and test developers with a quantitative method for determining qualitatively different groups of test takers.

An attractive feature of the approach used in this study is that intervals were first identified for the cutscores, rather than specific score points. The median score of students within these intervals were chosen as the cutscores. This practice is similar to the idea that "borderline" examinees should be considered in making standard setting decisions. However, other strategies for determining the cutscore within a score interval are possible. In any event, provision of a cutscore interval, rather than a single point, provides flexibility to standard setters who must consider political, resource, and other factors when deciding where to set the standards.

Future research should explore other methods for deriving cutscores from cluster analysis solutions. For example, given a score interval that seems to best separate clusters differing in proficiency, the score within this interval that is associated with the greatest test information (i.e., lowest conditional standard error of measurement) may be chosen as the cutscore. Thus, clustering approaches should be combined with emerging approaches for scaling and setting standards on educational tests to produce optimal results. In addition, the generalizability of the clustering approach needs to be further investigated with different types of tests and score distributions.

Another attractive feature of the clustering approach is the absence of a unidimensionality requirement, which has proved problematic in other standard setting procedures. Thus, the cluster-analytic approach appears promising for emerging performance assessments, which often depart from unidimensionality. An interesting side observation noted by Sireci (1995) is that by cluster analyzing examinees, groups of test takers with relative strengths and weaknesses across the different content areas may be observed, even when factor analysis of the test data indicate the test is measuring a unidimensional construct. Thus, cluster or factor analysis of *examinees*, rather than items, may provide new insights regarding test dimensionality. For all the clusters derived in this study, the cluster centroids revealed similar patterns across the four content areas comprising the test. This finding supports the conclusion that the test is unidimensional. However, in the Sireci (1995) study, different centroid patterns emerged revealing qualitatively different types of student-by-content area interactions. These types of interactions suggested multidimensionality, even though linear and non-linear factor analyses of the item response data indicated the test was essentially unidimensional.

Two features of the present study are critical for future applications of the procedure. The first is evaluating the stability of the cluster solution across samples; the second is external validation of the solutions. These two evaluations are necessary to ensure that the cluster solutions are stable and meaningful, rather than artifactual. Future applications with larger sample sizes should consider replicating the analyses over several samples.

The increasing consequences and stakes associated with educational and psychological tests demand increased rigor and accountability in the standard setting process. The results of this study indicate cluster analysis is a useful technique for psychometricians to add to their arsenal of standard setting procedures. Future research should evaluate the utility of these procedures with respect to other emerging procedures such as the dominant profile method (Putnam, Pence, & Jaeger, 1995). Cluster analysis is an under-utilized technique in educational measurement. Its application to the standard setting problem appears to be an area of great promise.

References

Anderberg, M. R. (1973). Cluster analysis for applications. New York: Academic Press.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. Applied Measurement in Education, 1, 215-222.

Cizek, G. J. (1993). Reconsidering standards and criteria. Journal of Educational Measurement, 30, 93-106.

Cizek, G. J. (1996). Setting passing scores. [An NCME instructional module]. Educational Measurement: Issues and Practice, 15 (2), 20-31.

Connecticut State Department of Education (1994). Connecticut mastery test technical report. Hartford, CT: Author.

Dalrymple-Alford, E. C. (1970). The measurement of clustering in free recall. Psychological Bulletin, 75, 32-34.

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. Psychological Assessment, 8, 360-362.

Friedman, H. P. & Rubin, J. (1967). On some invariant criteria for grouping data. Journal of the American Statistical Association, 62, 1159-1178.

Hubert L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin, 83, 1072-1080.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45, 325-342.

Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 46, 187-199.

Milligan, G. W. (1995). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. J. Hubert, & G. DeSoete (Eds.) Clustering and classification (pp. 345-375). River Edge, NJ: World Scientific Publishing.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50, 159-179.

Phillips, S. E. (1996). Legal defensibility of standards: Issues and policy perspectives. Educational Measurement: Issues and Practice, 15 (2), 5-19.

Plake, B. S. (1996). Setting performance standards for professional licensure and certification. Commissioned paper by the National Academy of Science's Committee on the Evaluation of National and State Assessments of Educational Progress.

Putnam, S. E., Pence, P., & Jaeger, R. M. (1995). A multi-stage dominant profile method for setting standards on complex performance assessments. Applied Measurement in Education, 8, 57-83.

Sireci, S. G. (1995, August). Using cluster analysis to solve the problem of standard setting. Paper presented at the annual meeting of the American Psychological Association, New York.

Sireci, S. G., & Robin, F. (1996, June). Setting passing scores on tests using cluster analysis. Paper presented at the annual conference of the Classification Society of North America, Amherst, MA.

Sneath, P. H. (1980). The risk of not recognizing from ordinations that clusters are distinct. Classification Society Bulletin, 4, 22-43.

SPSS, (1993). SPSS for Windows, version 6.0 (computer program). Chicago, Il:

SPSS, (1993). SPSS for Windows, version 6.0 (computer program). Chicago, Il: Author.

U.S. General Accounting Office, (1993). Educational achievement standards: NAGB's approach yields misleading interpretations (GAO/PEMD Publication No. 93-12). Washington, DC: Author.

Zieky, M. J. (1994, October). A historical perspective on setting standards. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Table 1

Recoding of Final Grades to Account for Educational Tracks

	Original Grade				
Recoded Grade	F	D	C	B	A
High Track Student	3	4	5	6	7
Typical Track* Student	2	3	4	5	6
Remedial Track Student	1	2	3	4	5

Note: Includes all students from the magnet school.

Table 2

C-indexes for Hierarchical Cluster Analyses

Solution	Sample A	Sample B
2-Clusters	3.8[1]	38.0
3-Clusters	4.5	1.9[1]
4-Clusters	3.8[1]	2.9
5-Clusters	4.6	3.2
6-Clusters	11.0	4.0

Note: The smallest C-index indicates the best cluster solution.

Table 3

C-indexes for K-means Cluster Analyses

Solution	Sample A	Sample B
2-Clusters	61.1	48.0
3-Clusters	18.6	7.5
4-Clusters	9.0	5.6[1]
5-Clusters	13.8	8.6
6-Clusters	7.8[1]	9.4

Note: The smallest C-index indicates the best cluster solution.

Table 4: Summary of K-means Cluster Analyses Using Content Subscores

Cluster Solution	Cluster/ Sample	Prop. in Cluster	Median Tst. Scr.	Min. Test Score	Max. Test Score	Concepts	Comp./ Estimt.	Probl. Solving	Meas./ Geom.
2	1A	.47	68	33	95	-.9	-.8	-.9	-.8
	2A	.53	122	87	142	.9	.9	.9	.8
	1B	.49	68	31	93	-.8	-.8	-.8	-.9
	2B	.51	117	87	143	.7	.7	.8	.7
3	1A	.31	59	33	75	-1.1	-1.1	-1.2	-1.0
	2A	.33	94	73	112	.0	.0	.0	.1
	3A	.36	128	107	142	1.1	1.1	1.1	1.1
	1B	.33	60	31	78	-1.1	-1.1	-1.1	-1.0
	2B	.34	91	72	111	.0	.0	.0	.1
	3B	.33	123	104	143	1.0	1.0	1.0	1.0
4	1A	.20	52	33	64	-1.4	-1.3	-1.4	-1.2
	2A	.25	76	63	93	-.5	-.5	-.5	-.6
	3A	.26	104	81	122	.4	.3	.4	.3
	4A	.29	130	116	142	1.2	1.2	1.2	1.2
	1B	.28	59	31	72	-1.1	-1.2	-1.2	-1.1
	2B	.25	82	68	99	-.3	-.3	-.3	-.4
	3B	.27	108	95	122	.5	.5	.6	.4
	4B	.21	131	117	143	1.2	1.2	1.1	1.2
5	1A	.16	49	33	60	-1.5	-1.4	-1.5	-1.2
	2A	.19	70	56	79	-.7	-.8	-.8	-.7
	3A	.17	87	78	98	-.2	-.2	-.1	-.3
	4A	.21	107	96	122	.5	.5	.5	.4
	5A	.28	131	116	142	1.2	1.3	1.2	1.2
	1B	.16	52	31	62	-1.4	-1.4	-1.4	-1.1
	2B	.22	69	58	84	-.7	-.7	-.6	-.9
	3B	.19	90	72	102	-.1	-.1	-.1	.0
	4B	.24	111	100	124	.7	.6	.7	.4
	5B	.19	132	117	143	1.2	1.2	1.2	1.3
6	1A	.11	46	33	55	-1.5	-1.5	-1.6	-1.4
	2A	.17	63	51	75	-1.1	-1.0	-1.0	-.7
	3A	.09	80	62	95	-.4	-.2	.0	.4
	4A	.15	90	73	105	-.3	-.4	-.4	-.8
	5A	.21	108	91	123	.6	.5	.5	.3
	6A	.27	131	116	142	1.2	1.2	1.2	1.2
	1B	.11	52	31	64	-1.3	-1.4	-1.4	-1.3
	2B	.19	68	54	85	-.8	-.9	-.7	-.6
	3B	.17	83	69	99	-.2	-.1	-.3	-.8
	4B	.11	98	85	116	.0	.0	.2	.6
	5B	.18	111	97	124	.7	.6	.7	.2
	6B	.20	131	115	143	1.2	1.1	1.1	1.3

Table 5: Within-Cluster Math Grade Mean Comparisons

Cluster Solution	Cluster Number	Sample	Mean Grade	All Pairwise Significant?
2	1	A	4.1	
	2	A	5.8	yes
	1	B	4.1	
	2	B	5.6	yes
3	1	A	3.9	
	2	A	4.9	
	3	A	6.1	yes
	1	B	3.8	
	2	B	4.9	
	3	B	5.9	yes
4	1	A	3.8	
	2	A	4.4	
	3	A	5.1	
	4	A	6.3	yes
	1	B	3.7	
	2	B	4.6	
	3	B	5.4	
	4	B	6.2	yes
5	1	A	3.8	
	2	A	4.1	
	3	A	4.7	
	4	A	5.2	
	5	A	6.3	no
	1	B	3.6	
	2	B	4.1	
	3	B	4.8	
	4	B	5.5	
	5	B	6.2	no
6	1	A	3.6	
	2	A	4.0	
	3	A	4.4	
	4	A	4.6	
	5	A	5.3	
	6	A	6.3	no
	1	B	3.6	
	2	B	4.0	
	3	B	4.7	
	4	B	4.9	
	5	B	5.5	
	6	B	6.2	no

Table 6

Cross-tabulation of Rescored Final Math Grades by Cluster

Sample A

(n=391, Spearman r=.69)

Cluster

Grade	Intervention	Proficient	Excellence
1	1		
2	10		
3	29	7	2
4	50	36	6
5	25	56	20
6	6	29	59
7		2	53

Sample B

(n=383, Spearman r=.69)

Cluster

Grade	Intervention	Proficient	Excellence
1	1		
2	13		
3	32	8	2
4	46	37	7
5	23	55	30
6	7	23	54
7		7	38

Table 7

Cutscores and Classification Percentages Resulting From Different Standard Setting Methods

Method	Intervention/ Proficient Cutscore	Proficient/ Excellence Cutscore	Percentage Classified as Intervention	Percentage Classified as Proficient	Percentage Classified as Excellence
Cluster-analytic	75	107	31.0	31.2	37.8
Locally-derived	71	112	27.2	40.2	32.6
Criterion-referenced	54	127	10.1	72.5	17.4

Figure 1

Histogram and Descriptive Statistics for Total Test Score

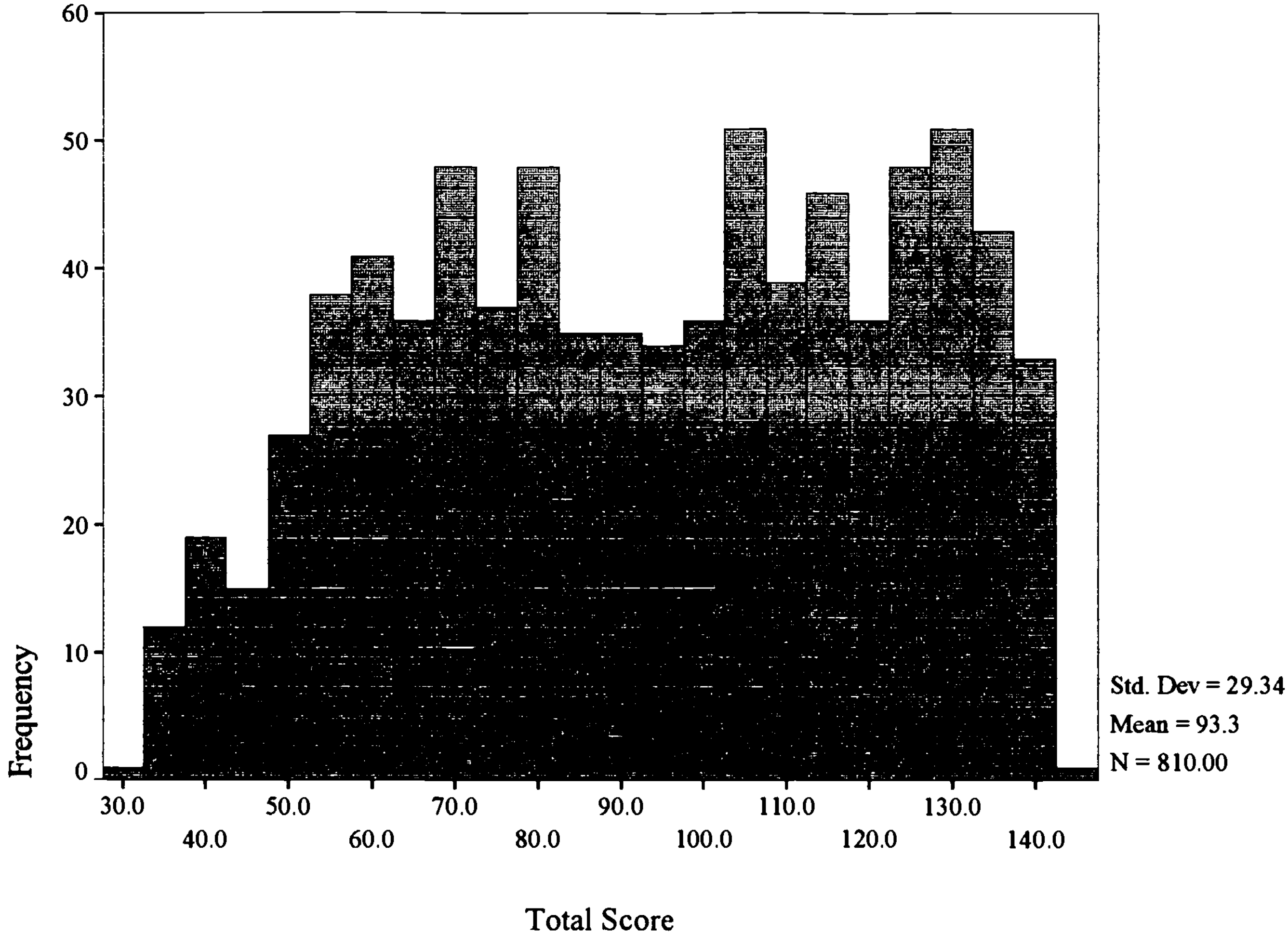
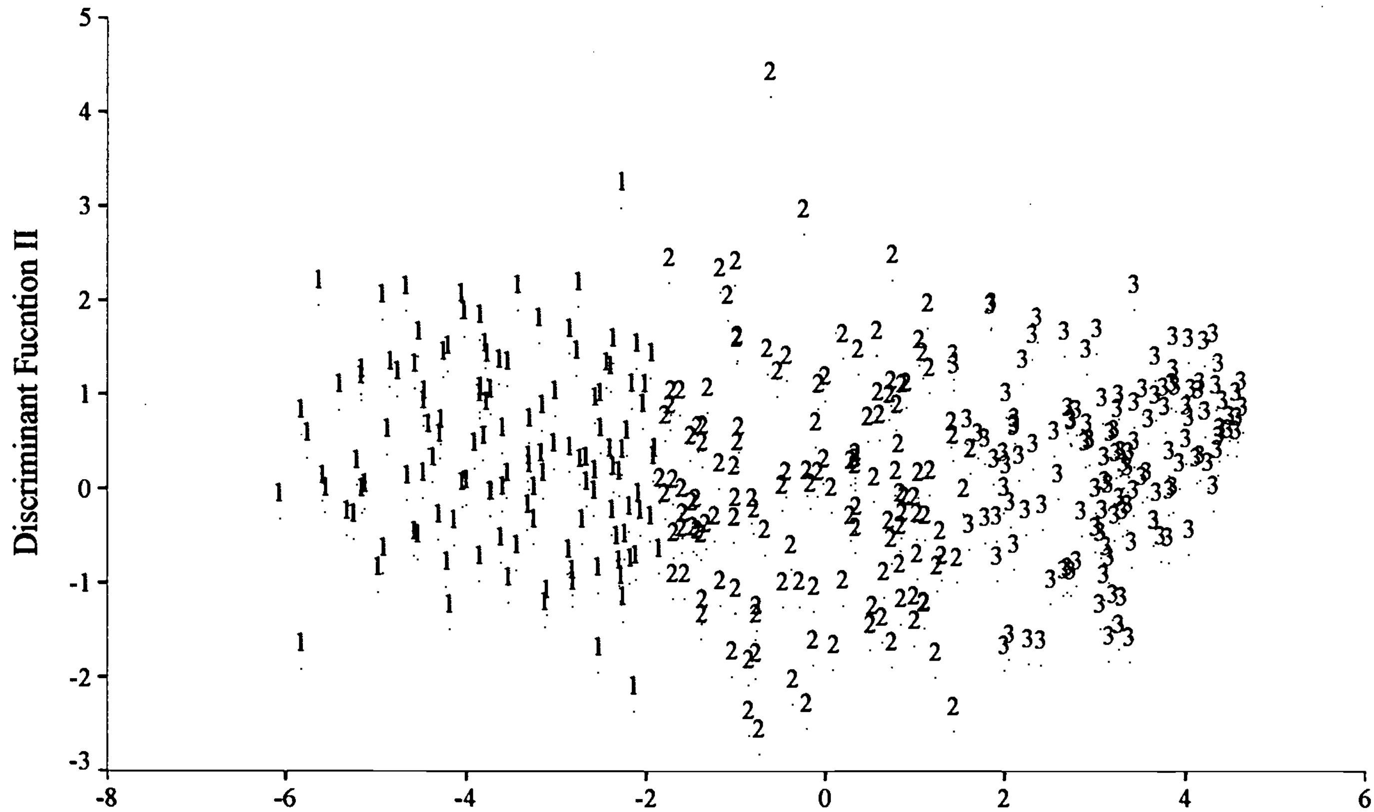


Figure 2:

Sample A Cluster Membership Plotted Using Discriminant Functions



AREA 1997

TM027724

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE
(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Using Cluster Analysis to Facilitate the Standard Setting Process

Author(s): Stephen G. Sireci, Frederic Robin, & Thomas Patelis

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

[✓] **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document [ ]

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
Sample
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: [signature]	Position: Assistant Professor
Printed Name: Stephen G. Sireci	Organization: University of Massachusetts
Address: School of Education 156 Hills South University of Massachusetts Amherst, MA 01003	Telephone Number: (413) 545-0564
	Date: 9/1/97

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 21, 1997

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1]If you are an AERA chair or discussant, please save this form for future use.