

Running Head: CLUSTER ANALYSIS AND STANDARD SETTING

Cluster Analysis and its Application

In

Standard Setting

Gregory S. Sadesky

University of Alberta

Introduction

Work by Stephen Sireci and his colleagues (Sireci, 1995; Sireci, Robin, & Patelis, 1999) demonstrate the utility of using cluster analysis to facilitate standard setting. Of particular interest was the capacity of cluster analysis methods to help isolate groups of examinees based on their performance on content-based subtests and varying response formats. These clusters were subsequently used to inform the setting of cut scores that maximally discriminated between the groups. However, as Kane (2001) asserts, standard setting is fundamentally a two component process: the setting of cut scores and the specification of performance standards that define what students can be expected to know and do at each achievement level. Thus, an equally important objective for cluster analysis and its application in standard setting is to help define those performance standards. The purpose of this paper is to show how information gained from cluster analysis can help bring about a clearer definition of those standards.

This paper is structured as follows. First, cluster analysis is briefly reviewed. Then, a review of Kane's (2001) analysis of the distinction between cut scores and performance standards will be undertaken. Next, the methods to establish the validity of a cluster solution on examinee data will be reviewed and it will be argued that these methods are most effective when they are informed by clearly defined performance standards. Sireci's influential work in this area will then be analyzed with respect to the extent to which defensible cut scores and performance standards are identified. After a discussion of future research possibilities, a review of the method in light of the criteria specified by Berk (1986) and Hambleton (2001) will be undertaken.

What is Cluster Analysis?

Cluster analysis is a set of statistical methods that group individual observations into classes (called clusters) on the basis of similarity. Of this set, the two most common cluster methods applied to standard setting are hierarchical and K-means cluster analysis.

Hierarchical cluster analysis (HCA). HCA comprises two separate methods, agglomerative and divisive. When using hierarchical agglomerative clustering, each individual observation is initially designated as a separate cluster. In a stepwise fashion, the most similar clusters are combined into larger units, ending when there exists one super-cluster containing all observations. In contrast, the divisive technique begins with the single super-cluster, and proceeds stepwise by dividing the cluster into its most dissimilar two parts. This process repeats, ending when there are n clusters, one for each observation. Hierarchical clustering can be used in standard setting to define a set of n cluster solutions and each solution can then be evaluated for its respective fit of the data.

K-means. In contrast to hierarchical clustering methods, K-means cluster analysis starts with the user identifying the number of clusters desired in the solution, and the *centroids* (cluster means) for each. An individual observation is compared with the values of each centroid and assigned to the cluster with which it is most similar. The value of each affected centroid is recalculated after each new assignment. The process is complete when, after a complete pass through the dataset, no re-assignments are made. The strength of this procedure for standard setting is that individual candidate cluster solutions can be compared on the basis of their fit to the observed data.

Mathematical Considerations

Similarity. Both of the above methods of cluster analysis use similarity between observations as the basis of categorization. Since all examinee data can be represented as *vectors* (one-dimensional arrays) similarity is defined geometrically. Although several alternatives exist for defining this similarity, the most commonly used is *Euclidian Distance*, defined between two vectors a and b each having n elements as:

$$= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2}.$$

This value is equivalent to the shortest distance between the two points in co-ordinate space.

What should compose the vector? A choice must be made of how to represent the examinee data in the cluster analysis. In the case where the test is composed of a number of different subscales or testlets, the vector representing student performance could consist of the total scores on each of these scales. If no explicit subscales exist, factor analysis could be performed on the data in advance of the cluster analysis and the scores on each factor for each examinee would serve as input to the cluster analysis. If item level information is desired from which to form the clusters, then scores on individual items could comprise the input vector. Last, if important differences exist in the distracters on multiple choice items that need to be considered, the actual student response to each item could be included.

How many clusters belong in the solution?

Both HCA and K-means cluster analysis can produce many solutions for a given problem. For example, HCA produces a set of cluster solutions whose number equals the

number of elements to be clustered. However, neither procedure provides information about the quality of the solutions that they identify. Therefore, some criteria must be available to provide selective support for some cluster solutions over others. There are many procedures to choose from (see e.g., Milligan, 1981) but a well-established rule-of-thumb is to choose the solution that maximizes the between cluster distance while minimizing within cluster distances. In the study by Sireci et al. (1999), a measure called the C-index (Dalrymple-Alford, 1970) is used to operationalize these criteria.

How Does Cluster Analysis Inform Standard Setting?

The result of a cluster analysis is a categorization of each observation in the sample. When observations are vectors of examinee responses to a test and the clusters represent levels of achievement, the cluster solution may be useful in setting cut scores. Two methods are particularly useful here, the *contrasting groups* and the *borderline group* methods (Livingstone & Zieky, 1982; 1989). If it is determined that a cluster in the solution represents a group of borderline or minimally competent examinees with respect to a given standard, a score that is typical of this group could be used as a cut score, for example, the median. This defines the borderline group method. The contrasting groups method could be employed when the clusters represent two groups of examinees on either side of a standard. In this case, the cut score could be set to minimize the classification errors in each of the two groups. For example, when a cluster contains only high scores and the examinees in an adjacent cluster contains lower scores, the boundary score that minimizes misclassifications between the two groups could be used as a cut score.¹

¹ Note that the transformation from cluster to cut scores using the two methods takes the clusters that are defined in terms of the examinee response vector and calculates a cut score in the total score or total ability metric -- a unidimensional index.

Following Kane (2001), setting cut scores is only one of two parts of the standard setting process. The other part involves defining a *performance standard* for each level of achievement. Such standards “provide qualitative descriptions of the intended distinctions between adjacent levels of performance...” “...in terms of what examinees [at each level] know and can do” (Kane, 2001, p. 55). In most standard setting situations, the performance standards are established first and cut scores are designed as an operational form of the standards. However, since a cluster solution depends only on the data and not the knowledge and skills that underlie them, the specification of performance standards would be relevant as part of the validity of the solution.

Kane argues that the specification of a performance standard is an integral part of validating any categorical judgments about the achievement of examinees. He states, “the aim of the validation effort is to provide convincing evidence that the cutscore does represent the intended performance standard and that this performance standard is appropriate, given the goals of the decision process” (Kane, 2001, p. 57). From this standpoint, to use cluster analysis to derive cut scores in the absence of a concise statement of intended performance standards is to frustrate the proper validation of the derived categorization. Thus, from Kane's standpoint, a cluster solution alone is insufficient to allow for the specification of valid achievement categories in data. In the next section, the types of evidence that can be gathered in support of the validity of a cluster solution and their dependence on clearly defined performance standards will be outlined.

The Validity of a Cluster Solution

Internal validity evidence. Examining the homogeneity of the cluster solution is an important step in validating it. Milligan (1981) argues that since any cluster analysis technique will impose clusters on even random data, the extent to which a given cluster solution exhibits both internal cohesion (within clusters) and external isolation (between clusters) determines whether natural categories exist in the data. When they do, this may be evidence that there are qualitative differences between levels of achievement in the underlying subject domain. In order to determine if this is the case, the basis upon which the clusters were formed can and should be examined. For example, Stout et al. (1996) used HCA to determine whether clustered subsets of test items were related after factoring out the total test score for each examinee. The defining features of these subsets were then examined and it was noted that they were based on identifiable factors, namely items all related to a single reading passage. In the standard setting context, when the resulting cluster solution relates to the achievement domain it can be interpreted as a source of confirmatory evidence for the validity of the solution. For example, examinees in a particular cluster were able to solve long division problems on a test while examinees in other clusters could not. Thus, when the defining features of cluster membership are part of the essential skills and knowledge in the domain being tested, those features could form part of the definition of the performance standard. When this is the case, the cluster solution itself supports the creation of valid performance standards and cut scores.

External validity evidence. The relationship of the categories created from cluster analysis to other criterion variables can also serve as evidence in evaluating the validity of the cluster solution. Criterion variables can be chosen because of their relationship

with the achievement domain, either convergent or divergent. The magnitude of the correlation between an ordered cluster solution and the criterion variable provides evidence for the validity of the categorization: the higher the absolute value of the correlation, the more compelling the evidence.

Critically, the criterion variable should be chosen so that the performance standards specified are inherent in the criterion. If appropriate evidence can be brought forth to show that they are, then the validity of the performance standards and cut scores can be appropriately evaluated through this correlation. When no such evidence is included, it is unclear how the relationship between criterion and categorization is to be interpreted. In the two studies to follow (Sireci, 1995; Sireci et al., 1999), student's final school grades are used to help validate the categories identified from cluster analysis.

The judgment of experts concerning the defining features of the categorization is a further source of external validity evidence. If experts in the content domain can identify the characteristics of identified clusters as being relevant to the construct underlying the test, this can lead to the specification of valid performance standards. Stevens and colleagues (e.g., Stevens et al., 1999; Vendlinski & Stevens, 2000) used a computational tool known as a self-organizing map (SOM) (Kohonen, 2001) to categorize examinee response vectors to complex reasoning problems. After categories were identified by the SOM, experts in the achievement domain examined the characteristics that defined them. The prototypes of each category that were rated highly by experts were characterized by deliberate and targeted movement through the problem space whereas prototypes rated as poor were described as unsystematic and random. Thus, experts corroborated the

categories detected by the SOM and thus provided key information in the specification of valid performance standards.

In summary, the above methods demonstrate the critical role for performance standards in establishing the validity of a cluster solution in standard setting. First, when data are naturally clustered and the clusters can be characterized by domain relevant factors, performance standards can be based on these factors. Second, when criterion variables correlate well with the cluster solution, performance standards can be constructed on which both the criterion variable and performance on the test critically depend. Last, subject matter experts can analyze clusters to identify the domain relevant elements that define them. When each of these methods is used, the validity of the interpretation of the cluster solution is enhanced. In the section to follow, prominent research conducted by Stephen Sireci and his colleagues is reviewed in this light.

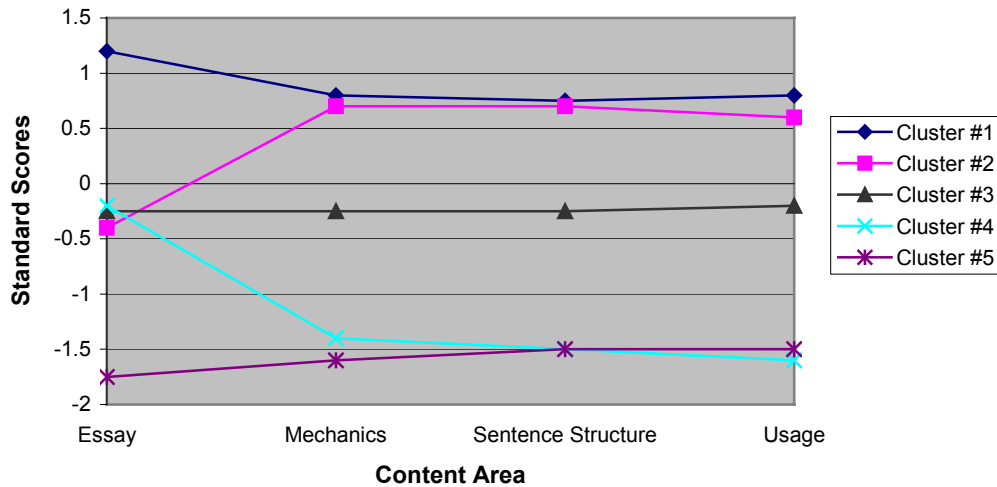
Using Cluster Analysis to Set Standards

In order to evaluate cluster analysis in the standard setting context, Sireci (1995) examined data from the General Educational Development (GED) test in writing skills. This test is administered to high school seniors to determine their English achievement level on four different subtests: essay writing, mechanics, sentence structure, and word usage. The test included two different item formats, written for the essay writing subtest and multiple-choice for the other three subtests. The scores on the test were being used to set standards for adults who were seeking equivalency credit in English towards a high school diploma.

The scores for each examinee on each of the four subtests were used as input for the cluster analysis. Using these data, Sireci determined that the optimal solution comprised

five clusters, as shown in Figure 1. This solution was identified as being optimal independently for each of two consecutive testing years.

Figure 1. GED Cluster Solution



The clusters identified reflected relative strengths and weaknesses of examinees on two factors, overall ability and item type. Specifically, clusters 1 and 2 both performed similarly well (above average) on all the three multiple-choice subtests, but were differentiated by their respective performance on essay writing. On this subtest, examinees in cluster 1 performed, on average, 1.5 standard deviations above the examinees in cluster 2. Similarly, examinees in cluster 4 outperformed those in the cluster 3 on the written essay subtest, but were indistinguishable as a group between the subtests of multiple-choice items. Cluster 3 comprised examinees that performed equally well on all four subtests, and marginally below average overall.

From a previous standard setting exercise, a cut score at the 30th percentile of the high school senior sample was designated. Sireci identified cluster 3 as likely containing the group of minimally-competent or borderline examinees, since the average

performance was equal across subtests, and slightly below average. Thus, he used the median total score of the students in this cluster as the cut score. This compared quite well with the previous score, differing by only 2 percentile points (32 vs. 30).

In analyzing the validity of the cluster solution for standard setting, it would be useful to know to what extent the identified clusters are homogeneous. No data were provided on this issue. However, Sireci did collect some criterion validity evidence, students' English grades. The correlation between the cluster and the grades was statistically significant, but low, 0.38. Finally, beyond the differences in performance on item type, no analysis was performed on the domain relevant features of each cluster.

Discussion. Through this simple application of cluster analysis to standard setting, Sireci demonstrated several strengths and weaknesses of the approach. First, he showed that the cluster solution could shed light on the nature of examinee similarities and differences, in this case, with respect to item types. Furthermore, he showed that this solution was stable, i.e., that the 5-cluster solution persisted across replications of the standard setting process. Last, he produced some criterion evidence for the validity of the categorization, namely a small positive correlation with school grades and a close match with the cut score previously set.

Clearly, however, these data do not reflect the full potential of cluster analysis in standard setting. First, it was not indicated what decisions were made with respect to the measure of distance or the type of cluster analysis used to determine the solution. Furthermore, we do not know to what extent the cluster solution was homogeneous, and apart from the differences in performance for particular examinees on different item types, the performance standards that defined the clusters remain largely unknown. Last,

the evidence of the validity of the cluster solution is not comprehensive, consisting of only one external criterion with which correlations were weak.

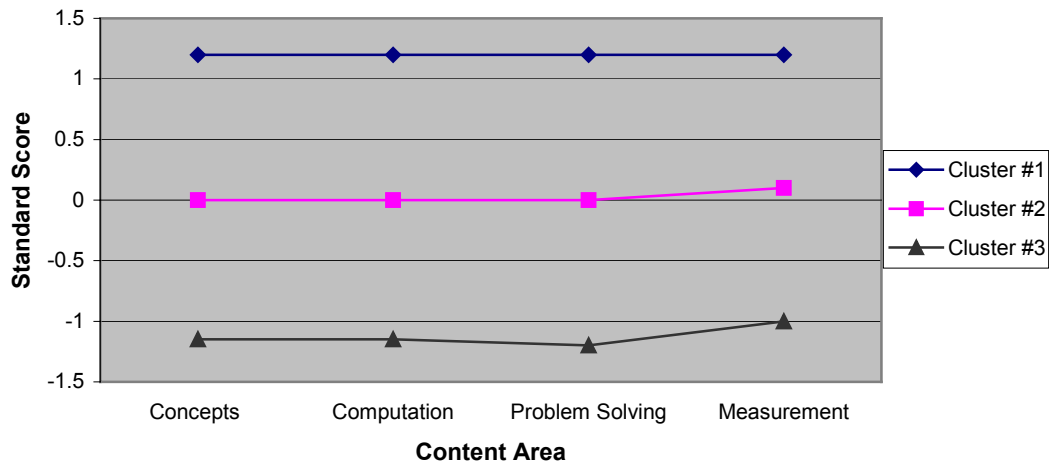
A subsequent study by Sireci et al. (1999) addressed some of the shortcomings brought to light by Sireci (1995). In this study, cluster analysis was used on data from the math portion of the Connecticut mastery testing (CMT) program for grade 7 students to validate an existing set of cut scores. These cut scores were initially set using a modified Angoff procedure. Like the previous study, students' grades were collected to provide evidence for the validity of the cluster solution.

All the items composing the test were multiple-choice from one of four topic areas: concepts, computation, problem solving, and measurement. Standard scores for each of these subscales were used as input to the cluster analysis. As a first step, Sireci et al. used agglomerative HCA to identify a set of possible cluster solutions for the data, choosing only the most homogeneous solutions identified using the C-index criteria (Hubert and Levin, 1976). A subset of the cluster solutions from HCA was identified as possible candidates for standard setting and was then regenerated using the K-means approach. Euclidian distance defined the similarity metric for each cluster analysis.

Two criteria were used to evaluate the best of the candidate cluster solutions: replicability across samples and significant pair-wise differences between all cluster centroids. Using these criteria, the three-cluster solution was identified as being optimal (see Figure 2). Interestingly, the previous standard setting using the modified Angoff procedure was designed to identify three categories of achievement (intervention, proficient, excellence) and thus could be directly compared with the present method. Further evidence for the validity of the three-cluster solution is the correlation between

cluster membership and school grade, in this case, 0.69. In comparison, the correlation between school grades and total test score was 0.72.

Figure 2. CMT Cluster Solution



Both the borderline group and contrasting groups procedures were used to set cut scores from these data. In the borderline group procedure, a cut score interval was defined as spanning the lowest test score from a given cluster and the highest score from the adjacent lower cluster. Following Livingston and Zieky (1989), the median test score was taken from examinees falling in this interval. The estimated cut score using this procedure was 75 for the intervention / proficient boundary and 107 for the proficient / excellence boundary.

Using the contrasting groups procedure, cut scores were set using logistic regression to minimize classification errors at the achievement level boundaries. Using this procedure, cut score estimates from consecutive years matched closely at both the intervention / proficient boundary (75.2 and 75.6) and the proficient / excellence boundary (110.5 and 109.4). This compares with 71 and 112 derived from the modified

Angoff procedure. The key change in the classification of examinees from the Angoff to cluster analysis method is an increase in the size of the intervention and excellence categories at the expense of the proficient.

Discussion. Sireci et al. more clearly identified the types of information that a cluster analysis can provide when setting standards and cut scores. First, they provided evidence that among a set of possible cluster solutions, three clusters proved the most homogeneous and therefore could be related to characteristic differences between the identified groups of examinees (Milligan, 1981). However, information about the precise nature of these differences with respect to performance standards is limited to the scores across each of the four subtests. Item level data may provide information that illuminates the substantive nature of the differences between clusters and thus could be used to define performance standards.

Criterion data similar to that provided in Sireci (1995) (i.e., school grades) provided the main part of the evidence used to externally validate the cluster solution. There was a significant correlation between the cluster solution and the test scores, thus lending limited evidence regarding the validity of the cluster solution. Last, the defining features of each cluster in terms of domain relevant features were not examined.

General Discussion

From the above two studies, Sireci has shown that cluster analysis can reveal characteristics of examinee data that are relevant to standard setting. In particular, he has demonstrated that examinees can be differentiated according to their scores on subscales and by extension, that examinees could be differentiated when systematic differences exist between them with any type of data provided (e.g., item, testlet). Further, he has

shown that the validity of a cluster solution can be evaluated both with respect to internal criteria (homogeneity within and isolation between clusters) and external criteria (correlation with other relevant measures). Last, he has demonstrated that cut scores can be set on the basis of the characteristics of the data and these can match well with cut scores set using other methods. Thus, the strength of cluster analysis in standard setting lies in its capacity to discover natural categories of examinees, set cut scores to best discriminate between them, and in a limited way, establish the relevance of these categories to achievement in a given domain.

Reflecting on Kane's view that cut scores and performance standards both must be clearly defined in order to validate any standard setting scheme, more effort needs to be directed towards creating performance descriptors on the basis of a cluster solution. Though cluster analysis is a promising technique in detecting qualitative differences in examinee achievement since it is attracted to solutions with natural groupings, there is no guarantee that the solution identified is relevant to achievement in the domain, or even that a solution identifies real clusters. In personal communication with Sireci (Dec. 2, 2002), he indicated the possibility that the cluster solution in Sireci et al. (1999) could be an artefact of an underlying normal distribution: one cluster was identified at the mode, and the other two above and below this point. In general, this misinterpretation is acutely more possible when the underlying test is essentially unidimensional. Clearly, if the underlying data are not naturally categorical, applying a statistical technique whose unique strength is in detecting categories will not confer any specific advantage over other techniques. Thus, in order for cluster analysis to be most useful to the process of setting standards and cut scores, clusters must exist in the data.

Under what conditions should these clusters be expected? If the underlying domain is naturally categorical, for example, mastery is achieved through stages, the distribution of scores may be composed of several *modes*, each composed of students having mastered all of the more basic stages, and none of the more advanced. The most basic example of a stage-like difference is that characterizing the transition between random and systematic responding. In this case, it would be expected that examinees that respond in a systematic way to a set of test items (i.e., that reflects understanding of the domain) would be more similar to each other than to those who respond randomly to those items and thus the higher achieving examinees would form a cluster.

Other cluster arrangements are possible. While low-ability examinees might respond to difficult items in a random manner, average-ability examinees might respond consistently in a way that reflects a misunderstanding about the item, perhaps through an over-generalization of an easier concept. If we suppose that there is only one type of misunderstanding possible, then the data underlying this pattern might have two distinct clusters: systematic responding with misunderstanding, and those having the correct understanding. Of course, in order to be able to detect such examinees, items and tests would have to be designed to incorporate misunderstandings.

The essential point is that cluster analysis can be used to make meaningful categorizations of performance standards and not only to set cut scores. That is, cluster analysis could provide more valid information about examinee achievement when it is used to determine performance standards at each level of achievement. By employing the methods described in this paper, determining these standards are acutely more possible and thus the validity of the categorization from cluster analysis is improved.

Evaluating Cluster Analysis as a Standard Setting Method- Berk's (1986) and
Hambleton's (2001) Criteria

Both Berk (1986) and Hambleton (2001) developed criteria that they argued could be used to evaluate the merits of any standard setting procedure. In this section, those criteria are applied to cluster analysis as a standard setting method. Scores on Berk's criteria are presented in Table 1, and scores on Hambleton's criteria are provided in Table 2.

Berk's (1986) Criteria - Technical Adequacy

1. *The method should yield appropriate classification information.* In the sense that the method can classify examinees according to competence this is accomplished. However, there is limited information about the *meaning* of such a categorization. Score: **3**

2. *The method should be sensitive to examinee performance.* Cluster solutions that do not reflect differences in ability between examinees are not solutions at all. A cluster solution does not guarantee relevance to examinee ability. What is needed is a substantive account of the differences in difficulty between clusters. Score: **2**

3. *The method should be sensitive to instruction or training.* Though in the research reviewed in the present paper, no data were presented on this issue, it could be expected that if items are designed to reflect developing competency, so would clusters derived from performance on those items. Score: **3**

4. *The method should be statistically sound.* Given that standard setting can be viewed as a form of statistical pattern classification, a method used to accomplish this pattern classification in general is appropriate. Score: **4**

5. *The method should identify the true standard.* This method is ambivalent with respect to the nature of the data to be clustered. True score or observed score, a cluster solution will be found. Score: 4

6. *The method should yield decision validity evidence.* Concern about validity is the Achilles heel of cluster analysis in standard setting. Information about performance standards would help address this, but there is no guarantee that a valid solution can be found by cluster analysis. Score: 2

Berk's (1986) Criteria - Practicability

7. *The method should be easy to implement.* This method is particularly easy to implement given that no judges are required at any stage of the procedure. Score: 4

8. *The method should be easy to compute.* Anyone with knowledge of cluster analysis procedures and access to software should be able to complete the process, though there are certain intricacies that may prevent novices in statistics from performing the analysis. Score: 3

9. *The method should be easy to interpret to laypeople.* Unfortunately, statistical solutions are often difficult for laypeople to interpret. However, if performance standards were also provided, the solution would be more interpretable. Score: 2

10. *The method should be credible to laypeople.* As with #9 above, statistical solutions can be seen to lack credibility. Score: 2

It should be noted that many of the low ratings of this procedure resulted from the difficulty in interpreting the cluster analysis substantively. An enhanced method that incorporated some or all of the methods described in the paper would improve these ratings.

Hambleton's (2001) Criteria

1. *Was consideration given to the groups who should be represented on the standard-setting panel and the proportion of the panel that each group should represent?*

For the Score, no such panel is utilized. Score: *N/A*

2. *Was the panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on the educational assessment?* Score: *N/A*

3. *Were two panels used to check the generalizability of the performance standards? Were subpanels within a panel formed to check the consistency of performance standards over independent groups?* Score: *N/A*,

4. *Were sufficient resources allocated to carry out the study properly?*
Insofar and as was conceived, the resource requirements of this method are small as compared with other methods. However, resources requirements would increase if content experts would be available to help determine the performance standards implied by the cluster solution. Score: **3**

5. *Was the performance standard-setting method field tested in preparation for its use in the standard-setting study, and revised accordingly?* The design of the study by Sireci et al. (1999) benefited from previous work by Sireci (1995) and modifications were made to procedure accordingly. Of course, enhancements can still be made. Score: **3**

6. *Was the standard-setting method appropriate for the particular educational assessment and was it described in detail?* It is unclear whether it should be expected that data underlying performance in mathematics achievement should be

categorical. However, if it is not, the method is no worse off than others. Detail was sufficient. Score: **3**

7. *Were panelists explained the purposes of the educational assessment and the uses of the test scores at the beginning of the standard-setting meeting? Were panelists exposed to the assessment itself and how it was scored?* Score: *N/A*

8. *Were the qualifications and other relevant demographic data about the panelists collected?* Score: *N/A*

9. *Were panellists administered the educational assessment, or at least a portion of it?* Score: *N/A*

10. *Were panelists suitably trained on the method to set performance standards?* Score: *N/A*

11. *Were descriptions of the performance categories clear to the extent that they were used effectively by panelists in the standard setting process?* Though no panel existed, performance categories were not defined in the current study. Score: **1**

12. *If an iterative process was used for discussing and reconciling rating differences, was the feedback to panelists clear, understandable, and useful?* Score: *N/A*

13. *Was the process itself conducted efficiently?* A clear procedure was outlined in the study and, due to the minimal resources required, was carried out efficiently. Score: **4**

14. *Were the panelists given the opportunity to 'ground' their ratings with performance data and how were the data used?* The entire procedure depended on performance data exclusively and could benefit from the opportunity to 'ground' the clusters with panelists' input. Score: **3**

15. *Were panelists provided consequential data (or impact data) to use in their deliberations and how did they use the information?* The cut scores derived from the cluster procedure were compared with those derived from a previous standard setting exercise. Though it is unclear what impact a large mismatch would have had on subsequent cluster analysis, such data were considered in evaluating the success of the procedure. Score: **3**

16. *Was the approach for arriving at final performance standards clearly described and appropriate?* No such standards were arrived upon. Score: **1**

17. *Was an evaluation of the process carried out by the panelists?* Score: **N/A**

18. *Was evidence compiled to support the validity of the performance standards?* No performance standards were specified, but evidence was compiled in support of the appropriateness of the cut scores. However, the relationship between the cut scores and any valid performance standard remains an open question. Score: **2**

19. *Was the full standard-setting process documented?* Current procedure was well documented although certain characteristics of the cluster solution were not fully explained, namely the extent to which the underlying data were categorical or continuous. Score: **3**

20. *Were effective steps taken to communicate the performance standards?* No performance standards were defined. Score: **1**

Conclusions

Cluster analysis is in its infancy as a technique applied to standard setting. What has been learned so far as a result of the work by Sireci, Stevens, and others is that the

power of the technique in detecting clusters can be of great benefit when the clusters relate directly to the standards of performance, but is of uncertain value when its use is limited to the setting of cut scores. Thus, the validation of any cluster solution is a critical step in drawing inferences about individual examinee achievement. The technique shows promise when combined with careful test construction and expert judgment to maximize the interpretability of cluster solutions. The determination of the necessary conditions to achieve this will await future research.

References

- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced test. *Review of Educational Research, 56*, 137-172.
- Dalrymple-Alford, E. C. (1970). The measurement of clustering in free recall. *Psychological Bulletin, 75*, 32-34.
- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 89 – 117). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin, 83*, 1072-1080
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 53 – 88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.). Berlin: Springer-Verlag.
- Livingston, S. A., & Zieky, M. J. (1982). *Passing Scores*. Princeton, NJ: Educational Testing Service.
- Livingston, S. A., & Zieky, M. J. (1989). A Comparative Study of Standard-Setting Methods. *Applied Measurement in Education, 2*(2), 121-141.
- Milligan, G. W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika, 46*(2), 187-199.

- Sireci, S. G. (1995, August). *Using cluster analysis to solve the problem of standard setting*. Paper presented at the meeting of the American Psychological Association, New York.
- Sireci, S. G., Robin, F. R., & Patelis, T. (1999). Using Cluster Analysis to Facilitate Standard Setting. *Applied Measurement in Education*, 12(3), 301-323.
- Sireci, S. G. (2001). Standard setting using cluster analysis. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 339-354). Mahwah: Lawrence Erlbaum.
- Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, J., & Clyman, S. (1999). Artificial neural network-based performance assessments. *Computers in Human Behavior*, 15, 295-313.
- Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20(4), 331-354.
- Vendlinski, T., & Stevens, R. (2000). The use of artificial neural nets (ANN) to help evaluate student problem solving strategies. Paper presented at the Fourth International Conference of the Learning Sciences.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, 39(4), 291-309.

Table 1.

Ratings of Cluster Analysis as a Standard Setting Procedure: Berk's (1986) Criteria

Criteria	Score			
	1	2	3	4
<u>Technical Adequacy</u>				
1. Appropriate classification information			X	
2. Sensitive to examinee performance		X		
3. Sensitive to training			X	
4. Statistically sound				X
5. Identifies the true standard				X
6. Yields decision validity evidence		X		
<u>Practicability</u>				
7. Easy to implement				X
8. Easy to compute			X	
9. Easy to interpret to laypeople		X		
10. Credible to laypeople		X		

Table 2.

Ratings of Cluster Analysis as a Standard Setting Procedure: Hambleton's (2001)

Criteria

Criteria	Score				
	N/A	1	2	3	4
1. Panel representation	X				
2. Size of panel	X				
3. Two panels - generalizability	X				
4. Sufficient resources					X
5. Field-tested				X	
6. Appropriateness of method				X	
7. Purposes explained to panel	X				
8. Panelist demographics collected	X				
9. Test administered to panel	X				
10. Panelists well-trained	X				
11. Clear performance categories		X			
12. Clear iterative process	X				
13. Efficiency of process					X
14. Performance data provided				X	
15. Impact data provided				X	
16. Determining performance standards		X			
17. Evaluation of process	X				
18. Validity of performance standards			X		
19. Process documented				X	
20. Steps to communicate standards		X			