# CLUSTER ANALYSIS – A STANDARD SETTING TECHNIQUE IN MEASUREMENT AND TESTING

**Muhammad Naveed KHALID**
University of Cambridge ESOL Examinations, UK

**E-mail:** Khalid.M@cambridgeesol.org

**Abstract:** *Standard setting plays an important role in educational and psychological testing. This paper is focused on standard setting using 'cluster analysis' technique. Cluster analysis is a statistical procedure for forming homogenous groups of subjects (examinees). It explores the process of doing cluster analysis and its types are – K-Means and Hierarchical clustering. In the hierarchical cluster analysis, all objects are initially being considered to be a unique cluster. The analysis proceeds sequentially by merging clusters together one step at a time until all objects are merged into a single cluster. In the K-Means cluster analysis, the number of clusters into which the objects which will be portioned is specified initially. The K-means algorithm then establishes the centers of each cluster which are represented by a vector of means (called the cluster centroid) corresponding to the variables used to cluster subjects. The procedure was applied to an achievement test in science. A five cluster solution best separated the examinees according to their proficiency skills. The study concludes that cluster analysis has an edge over other techniques in regard to reducing subjectivity based on expert ratings of items and applicability to performance-based assessments. It does not remove subjectivity from the standard setting process, but does provide subject-matter experts and test developers with a quantitative method for determining different groups of test takers.*

**Key words:** *Cluster analysis; Standard Setting; K-Means Clustering; Hierarchical Clustering*

## 1. Introduction

Standard setting is an important and perennial problem in educational and psychological testing. It plays a significant role in teaching profession in selecting the most competent examinees for various purposes. It has also become very important because of the legal and political implications of having crude selection criteria based on non objective standard setting. Over the years many methods and techniques have been evolved for standard setting. Many of these techniques share some common features with each other; others differ to a large or small extent. One of such techniques for developing standard setting in science is 'cluster analysis'. This technique is aimed at clustering examinees with similar profiles such that the task of standard setting becomes easier. This technique provides standard setters with options and perspectives than other techniques do not. It is more embedded in statistical procedures than most other techniques which make use of subjective judgments. Though cluster analysis is a technique that can be reinforced by using external validity criteria, it is also possible to execute the full process using statistical procedures alone without resorting to external subjective opinion. This makes it a powerful tool for standard setters.

Cluster analysis, unlike many other methods, also provides standard setters with multiple possibilities for standard setting and can give additional insights into the multidimensional competencies of examinees. Though cluster analysis may be used on its own for the purpose of standard setting it is perhaps better suited to enter the realm of standard setting as a supporting tool for other standard setting methods, like the Reckase charts[1] and in the future it may be better to expand efforts on consolidating such a position for it. Availability of methods like Cluster Analysis and Reckase Charts as supporting tools can give different perspectives into test taker performance which can assist in strengthening the process of standard setting. Thus the cluster analysis technique is a promising research avenue that can elevate the science of standard setting to the next level, either on its own or at-least in conjunction with other methods as a valuable support tool as discussed in this paper.

### 1.1. Concept of Cluster Analysis

Cluster Analysis is a statistical procedure for forming groups of similar objects. It finds a broad range of application in many fields apart from standard setting exercises in the field of education. For example, in medicine, cluster analysis is used to identify diseases and their stages: by examining patients who are diagnosed as depressed, one can find if there are several distinct sub-groups of patients with different types of depression. In marketing, cluster analysis is used to identify people with similar buying habits; by examining these characteristics one may be able to target future marketing strategies more efficiently. In the field of education cluster analysis is a relatively new technique for standards setting purposes. It is currently still being developed and seems to hold a lot of promise for the future.

Traditional standard setting-setting methods have been criticized due to reliance on untested subjective judgment, lack of demonstrated reliability and lack of external validation. Cluster Analysis builds on the strengths of other standard setting methods and addresses some of their weaknesses. In particular the method includes replication and the use of external evidence of validity while relying less on subjective judgment.

### 1.2. Current Standard Setting Methods

Jaeger (1989; 1995) classified standard setting methods as either *test* centered or *examinee* centered. Test centered methods involve the use of expert panelists to scrutinize items comprising the test and to make judgments regarding the probable levels of performance that borderline or marginally proficient test takers will exhibit on the items. The most popular test-centered method is the Angoff method and its modifications. For items that are scored dichotomously, the panelists in an Angoff study estimates the probability that a borderline examinee will answer an item correctly. For items scored polytomously, the panelists estimate the expected score of a borderline candidate on the item. Cutscores are set on the test score scale by summing the item probabilities or expected item scores for each judge and then averaging these sums across panelists or taking the median score. There are two primary criticisms of test-centered methods for standard setting. First, the cognitive task presented to the panelists is complex, and it is difficult to provide evidence that they understand the task or complete it as desired (Angoff, 1988; Cizek, 1996). Part of this difficulty results from the notion of a borderline test taker; it may be difficult for panelists to clearly envision the knowledge and skills characterizing this test taker and to compare these levels of knowledge and skills to those required for success on numerous test items. A second criticism of test-centered procedures is that the resulting passing standard may change if a different group of panelists is used (Angoff, 1988; Cizek, 2001). Though some change is acceptable because of the element of subjectivity and the fact that there are no golden standards, it is a good exercise to limit the subjective element in the standard setting process. Significant discrepancies due to subjective opinions of different panelists can be a serious threat to the defensibility of cut-scores.

Examinee-centered standard-setting methods use subject-matter experts to evaluate examinees rather than items. One such approach, the borderline group method, uses experts to select a group of test takers who are considered marginally proficient (i.e., who possess just enough knowledge and skills to be classified into a particular category). The median test score for this borderline group is then used as the relevant cut-score. To implement the borderline group method, the borderline test takers must be selected using criteria other than test performance. This requirement poses problems because there is no direct way to determine borderline proficiency. Thus, the same types of false-positive and false-negative classification errors associated with standard setting in general apply to the assignment of test takers to the borderline group. In addition, the cut-scores derived in this fashion would fluctuate directly with the (likely to be unknown) sampling variability over potential borderline groups.

Another popular examinee-centered method is the contrasting groups' method. In this method, experts select two groups of test takers, one considered to be above the relevant standard and one considered to be below this standard. The test scores that result in the fewest false-positive (e.g. passing a below-standard student) and false-negative (e.g. failing an above-standard student) misclassifications are selected as the passing score. Though it is easier to identify above standard and below standard groups than to identify borderline groups, in many cases identification of contrasting groups is not easy to validate. The resources required for identifying and testing above and below standard students are much larger compared with the borderline method. Overall, both methods share many short-comings like unknown sampling variability across examinee groups, classification errors in assigning examinees to groups, and practical constraints.

A review of traditional standard-setting methods reveals that, although each method has theoretical appeal, all are subject to significant limitations. Several researchers have suggested guidelines or standards for evaluating standard-setting studies (Kane, 1994b; Van der Linden, 1994; Cizek, 1996) Furthermore, it suggests that standard-setting studies can be improved by: a) including replications of the procedure to evaluate the consistency of the resulting standards; b) incorporating validity checks on the resulting standard (e.g. convergent validity with external criteria); and c) using more than one standard-setting method.

### 1.3. Process of Cluster Analysis

The standard setting problem is essentially a classification problem (Sireci, 1995). When standards are set on a test, the purpose is to classify each test taker into one or more groups, such that test takers with abilities close to each other should be separated from those test takers with abilities that are different such that all test takers can be classified into categories or groups with similar ability levels. Cluster analysis does exactly that, i.e. groups test takers into homogeneous clusters with respect to the proficiency measured. Each cluster is comprised examinees that are highly similar in proficiency. These clusters can then be ordered in a manner congruent with the groupings defined by the standard-setting problem.

Cluster analysis can force discrete decisions on a continuous scale. When cut-scores are used to classify test takers into one or more groups, 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 but how to best partition test takers into the desired number of (discrete) groups motivated by the testing purpose. However the strength of cluster analysis can also be its shortcoming. The reason is that clustering procedures cluster the data regardless of whether truly different groups of examinees are present or not. Secondly, because it focuses on analysis of test response data, no standards can be set higher or lower than the test takers actually perform. The procedure derives standards based on what specific groups of test takers have done, rather than according to what they should have done. 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 would exhibit expected standards of performance. In this regard, we would like to quote some remarks of researchers who have analyzed cluster analysis results:

> *We applied the procedure to a state-wide mathematics proficiency test .The standards developed from cluster analysis were compared with those established at the local level and with those derived from a more traditional borderline and contrasting groups analysis. We observed relative congruence across the local cut score and those derived using cluster analysis, and we observed similar correlations among the resulting proficiency groupings and course grades. The results of the more traditional borderline and contrasting groups analysis were less favorable. We conclude that cluster analysis appears useful for helping set standards on educational tests (Sireci, 1999).*

### 1.4. Types of Cluster Analysis

There are two sorts of cluster analysis that can be used to form clusters. The first is called hierarchical cluster analysis and the second is called the K-Means cluster analysis. In hierarchical cluster analysis, all objects are initially being considered to be unique clusters. The analysis proceeds sequentially by merging clusters together one step at a time until all objects are merged into a single cluster. A “N-cluster” solution is, however of no practical use. The work for the standard setter is to determine the cluster solution in between these two extremes at which truly different clusters are merged together. The cluster solution preceding that point represents the best clustering of the data. The standard setter can make use of both internal and external criteria to help determine the optimal clustering solution. A severe limitation of this form of clustering is that once test takers are merged into a cluster, they are stuck for the remainder of the analysis, even if a rearrangement of test takers across clusters may improve the solution. Also this method is not suitable for large data sets due to the extremely large number of within and cross cluster comparisons that need to be made at each stage of analysis. However, in most educational standard setting exercises the goal is not to uncover the “true” cluster structure of the data but to identify the optimal partitioning of the examinee population that best corresponds to a stated number of groupings. Thus when the number of clusters in which examinees are to be partitioned is known at the start as in most educational instances the K-means clustering can be used. However, experienced examiners can use the hierarchical clustering as a preliminary to K-means clustering to have an estimate of how many real clusters there are that may then be specified in the K-means analysis.

In hierarchical clustering, clusters are formed by grouping cases into bigger and bigger clusters until all clusters are members of a single cluster. Before the analysis begins all cases are considered separate clusters: there are as many clusters as there are cases. At the first step, two of the cases are combined into a single cluster. At the second step either a third case is added to the existing cluster of two cases or two other cases are merged into a new cluster. At every step, either individual cases are added to the existing cluster or two new cases are merger into a new cluster. However once a cluster is formed it cannot be split. There are many criteria for deciding which cases or clusters should be combined at each step. A common method is the single linkage method; the first two cases combined are those that have the smallest distance between them. The distance between the new cluster and individual cases is then computed as the minimum distance between an individual case and a case in the cluster. The distance between cases that have not been joined do not change. At every step, the distance between two clusters is the distance between their two closest points. Another commonly used technique is called the complete linkage or the furthest neighbor technique. In this method, the distance between two clusters is calculated as the distance between their two furthest points. Yet another method is the average linkage between groups method, often called UPGMA which defines the distance between two clusters as the average of the distances between all pairs of cases in which one member of the pair is from each of the clusters. This differs from the other linkage methods in that it uses information about all pairs of distances, not just nearest or the furthest. Another method is the centroid method which calculates the distance between two clusters as the distances between their sums for all of the variables. In the centroid method, the centroid of a merged cluster is a weighted combination of the centroids of the two individual clusters, where the weights are proportional to the size of the clusters. In the median method the two clusters being combined are weighted equally in the computation of a centroid, regardless of the number of cases in each. This allows small groups to have equal effect on the characterization of larger clusters into which they are merged. When similarity measures are used, the criterion for combining is reversed, i.e. the clusters with large similarity based measures are merged.

Once the distance matrix between all cases and clusters has been calculated the actual formation of clusters commences which can be seen on an *icicle plot* or a *dendogram*. Both are graphical representations of the output. Commonly an *icicle plot* is used. An *icicle plot* is a graphical representation in which the clustering steps are shown on the vertical axis against the cases being clustered on the horizontal axis. The number of clustering steps is equal to the number of cases and at each step one case or cluster is combined with another case or cluster. Thus in step 1, there are as many clusters as cases and at every step the number of individual cases reduces by 1 until in the last step all cases have been merged into one cluster. The challenge for the examiner is to identify how many real clusters are there based on the results of the hierarchical analysis shown on the *icicle plot* or the *dendogram* etc.

In K-means clustering, the number of clusters into which the objects which will be portioned is specified initially .The K-means algorithm then establishes the centers of each cluster which are represented by a vector of means (called the cluster centroid) corresponding to the variables used to cluster test takers. For example, if test takers are being clustered based on their performance on four different sections of a test, the four means on each test section determine the centroid of a cluster, where the means are calculated using only those test takers in that cluster. The number of means constituting each centroid is equal to the number of variable used to cluster the objects. The number is denoted by K, hence the name K-means clustering. This type of scaling has two obvious differences from traditional psychometric scaling. First, the distance among test takers is not determined from a single mean but rather from a vector of means. Second, instead of test takers being placed on a continuous scale, they are placed into one of a discrete number of clusters. These clusters can be used to inform the standard-setting process by relating the examinee clusters to the proficiency groupings invoked by the standard setting and test development processes.

The typical K-means algorithm begins by searching through the data to find the Q test takers that are most different from one another with respect to the clustering variables e.g. sub-scores on the test, where Q represents the number of clusters specified in advance by the researcher. At this point, the K scores for these test takers are used as cluster centroids.

The K-means algorithm is iterative: each test taker is assigned to a cluster by computing the distance between the test taker and each cluster centroid and assigned to the cluster whose centroid it is closest to. Once all test takers have been assigned to the initial clusters, the cluster means are recomputed as an average of all cluster members and the clustering exercise is repeated. Some test takers are placed in a different cluster after every iteration. The iterations carry on until there is no test-taker movement across clusters. At this point the clusters are said to have stabilized and iterations finish. The resulting clusters are the final clusters; their membership represents the result of the clustering exercise.

### 1.5. Basic Steps in Cluster Analysis

Three main decisions need to be made in order to perform a cluster analysis on a set of data. The first is the selection of variables. This is a very crucial step. If important variables are excluded, poor or misleading findings may result. The variables chosen should be such that they cover the whole range of important factors that cause similarities or dis-similarities between the items. There are at-least three options for selecting the variables to be used for clustering the test takers: 1) use all individual items comprising the test, or 2) use orthogonal factor scores obtained from item level factor analysis, or 3) use sub scores derived from items comprising the major area of the test. The second decision is to look into 'how alike are the cases'? In cluster analysis, items are clustered on the basis of their nearness or closeness to each other. The nearness or closeness is measured in terms of their distance from each other. A commonly used index for distance between items is the either the Euclidean distance or squared Euclidean distance, which is the sum of the squared differences over all of the variables.

Euclidean distance (x,y) = $\{\sum_i (x_i - y_i)^2 \}^{½}$

Squared Euclidean distance (x,y) = $\sum_i (x_i - y_i)^2$

The third decision is regarding the criteria for combining clusters. There are many criteria for deciding which clusters or cases should be combined. All criteria are based on a matrix of either distances or similarities between pairs of cases. Often it is sum of Euclidean distances of the items from the vector of means of the clusters (centroids) which determine the placement of an item in any cluster.

## 2. Methodology

In order to render a judgment on whether cluster analysis should be used or not, we first intended to carry out a practical test of cluster analysis on a set of pre-marked data in order to discuss the results of cluster analysis in light of real evidence. Unfortunately, as we were unable to get hold of real score cards where grades of test takers were listed next to their test marks. It would have been interesting to compare the grades suggested by cluster analysis with the actual grades of the test takers. Nevertheless we would like to demonstrate the results of a cluster analysis carried out on a set of non-graded data and explain the result of the clustering exercise and relevant statistical information.

### 2.1. Data

We have done cluster analysis on the data set which consisted of 60 dichotomous items marked 1 or 0 and two polytomous items. The sample size for this study was 3000.

### 2.2. Defining Variables

The first step in the analysis was to define the variables. Given the large number of items comprising the test and the unknown possibility of inter-correlation among the content areas, we decided to use the method based on content areas sub-scores. Sub-scores for each of the content areas defined in the test were used as the input variables for cluster analysis.

On the basis of the test data available, it seemed best to partition the test into five content areas. The two polytomous items were left as they were but we decided to group the 60 dichotomous items into three groups of 20 items each as the data file suggested that there test content consisting of the dichotomous part consisted of three different sort of test areas of 20 questions each. The sub-scores for students in the dichotomous area were computed by summing their item scores within each content area. So we ended up with five variables: three for the 3 sub-sections of the dichotomous part and two for the 2 polytomous items. The next step was to decide if we wanted to standardize the content area sub scores prior to clustering to account for differences in the raw score scales due to any differences in the number of items in the content area .The number of items in each content area of the dichotomous section were equal i.e. 20 but the raw scores scale in the 2 polytomous items were lower. They were marked on a scale of 10 which was half of the scale in other sub-categories i.e. 20. For analysis we assumed that each content area was equally important and was supposed to have an equal bearing on the final grade. Thus it was needed to rescale the content area sub scores and bring them at par with each other so that they have an equal effect on the measurement of distances during cluster formation. It was decided to transform the polytomous items scale by doubling all the item scores in the polytomous content area to bring it at par with the dichotomous content areas scale. Thus each content area was now represented on a scale of 1 to 20.

The plan was: a) to perform a Hierarchical Cluster Analysis on the data file; b) to perform a K-means Analysis on the data file; and c) compare results of a K-Means and Hierarchical Analysis and suggest ways for improvement.

## 3. Discussion of Results

### 3.1. Hierarchical Analysis

Hierarchical cluster analysis was performed on the data set. The analysis suggested that a minimum of five clusters should be used for grouping the examinees. A large difference was found in the 'coefficients' column in the attached agglomeration schedule between a four cluster solution and a five cluster solution. The column labeled 'co-efficients' represents the distance between two combining clusters. By examining these values we got an idea about how unlike the clusters being combined are: small co-efficients indicate that fairly homogenous clusters are being merged while large co-effecients indicate that clusters containing quite dissimilar members are being combined. These coefficients can be used as guidance in deciding how many clusters are needed to represent the data. It is best to stop further clustering as soon as the increase between two adjacent steps becomes large. In our case, there was a significant increase of around 34 between the five cluster solution steps.

### 3.2. K-Means Analysis

The researchers decided to select a number of clusters suggested by the hierarchical cluster analysis which suggested that at-least 5 different clusters should be there. Thus all test takers in the K-means analysis were grouped in each of the 5 levels which they are closest to.

The cluster centroids for each cluster were determined by the K-means algorithm. It selected the N number of students (where N is the number of specified clusters) whose scores were most different from each other. After that using the Euclidean distance formula, the K-means algorithm placed the rest of the students in their respective clusters after calculating their distances from the K-means centroids. The process was iterative and carried on until there was no shifting of test takers across clusters, i.e. stability was achieved. For the given data set in the SPSS file and the number of clusters specified as 5, the results of the K-means clustering can be seen in the SPSS output file shown in Table I.

**Table I.** Initial Cluster Centers

| | Cluster | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| VAR0001S | 12.00 | 16.00 | 4.00 | 4.00 | 12.00 |
| VAR0002S | 12.00 | 14.00 | 2.00 | 2.00 | 4.00 |
| VAR00003 | 13.00 | 15.00 | 16.00 | 5.00 | 4.00 |
| VAR00004 | 6.00 | 20.00 | 19.00 | 5.00 | 14.00 |
| VAR00005 | 14.00 | 19.00 | 14.00 | 4.00 | 10.00 |

**Table II.** Iteration History

| Iteration | Change in Cluster Centers | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 6.942 | 7.296 | 6.462 | 5.910 | 5.605 |
| 2 | .807 | 1.139 | .780 | 1.691 | .570 |
| 3 | .371 | .000 | .209 | .476 | .834 |
| 4 | .172 | .199 | .306 | .313 | .594 |
| 5 | .395 | .237 | .000 | .200 | .187 |
| 6 | .202 | .176 | .000 | .294 | .419 |
| 7 | .138 | .000 | .000 | .172 | .000 |
| 8 | .122 | .000 | .000 | .202 | .142 |
| 9 | .120 | .000 | .000 | .000 | .129 |
| 10 | .000 | .000 | .000 | .000 | .000 |

Tables I-IV represent the output of a K-Means clustering exercise for the data set specified before. The numbers of clusters specified were five. Table I shows the initial cluster centers selected by the K-Means algorithm. Table II shows the iteration history from which it can be seen that after 10 iterations the all clusters were stabilized into the final form. A convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any cluster is .000. The current iteration is 10. The minimum distance between initial centers is 13.638.

**Table III.** Final Cluster Centers

| | Cluster | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| VAR0001S | 9.16 | 13.28 | 7.53 | 6.28 | 11.19 |
| VAR0002S | 5.96 | 7.54 | 3.84 | 4.61 | 5.43 |
| VAR00003 | 12.33 | 15.72 | 14.39 | 7.06 | 9.55 |
| VAR00004 | 10.29 | 16.33 | 15.39 | 9.36 | 14.38 |
| VAR00005 | 12.93 | 17.44 | 16.34 | 8.94 | 12.29 |

Table III shows the final cluster centers and Table IV shows the distance between the final cluster centers. Table V shows the cluster membership. As can be seen from the table the cluster membership is fairly even i.e. the examinees are fairly evenly spread across the five clusters, which is a desirable feature of an exam.

**Table IV.** Distance between Final Cluster Centers

| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | 9.370 | 7.004 | 7.397 | 5.417 |
| 2 | 9.370 | | 7.115 | 15.918 | 8.788 |
| 3 | 7.004 | 7.115 | | 12.130 | 7.545 |
| 4 | 7.397 | 15.918 | 12.130 | | 8.208 |
| 5 | 5.417 | 8.788 | 7.545 | 8.208 | |

**Table V.** Number of Cases in each Cluster

| | | |
|---|---|---|
| Cluster | 1 | 45.000 |
| | 2 | 39.000 |
| | 3 | 38.000 |
| | 4 | 36.000 |
| | 5 | 42.000 |
| Valid | | 200.000 |
| Missing | | .000 |

Clusters can now be ordered into a hierarchal order by content experts if certain content areas are to be given priority over others or simply by summing the means of each final cluster centroid and then placing them in ascending order according to their net total scores with the highest number representing the highest cluster. After the clusters have been aligned in a hierarchical order, the cut scores can then be set. One way could be to set the mean scores of clusters, i.e. cluster centroids as the cut-off scores. Another way could be to identify the overlapping regions between clusters and then take the mean score of the overlapping regions to be the cut scores. Yet another way that could better determine the middle point of the overlapping region would be to take the median score of the overlapping region as the cut-score. The median method reduces the effects of any large variances in test scores of individual test takers on the whole group of test takers in the region under study. With this particular method borderline students can be better identified. Border-line students would be those who lie in the overlapping regions and would barely pass or fail depending upon their position with respect to the mean score of the over lapping region. It would be interesting to see how much variance exists between taking the median cluster scores or median of overlapping regions as the cut scores?

It is also possible to carry out other statistical procedures on the cluster items to determine the variance of variables within and across different clusters. Using this, we can observe how student response to certain item sets i.e. the variables varies across clusters. A high ratio of inter-cluster vs. intra-cluster would mean the variable varying significantly across clusters. This can give an insight into how clusters differ from each other. To do this a one way ANOVA is done on the data set as shown in the table VI.

**Table VI.** Inter-cluster and Intra-cluster differences through ANOVA

| | Cluster | | Error | | | |
|---|---|---|---|---|---|---|
| | Mean Square | df | Mean Square | df | F | Sig. |
| VAR0001S | 301.075 | 4 | 5.810 | 195 | 51.819 | .000 |
| VAR0002S | 76.121 | 4 | 6.418 | 195 | 11.861 | .000 |
| VAR00003 | 472.557 | 4 | 5.576 | 195 | 84.752 | .000 |
| VAR00004 | 383.855 | 4 | 5.288 | 195 | 72.587 | .000 |
| VAR00005 | 433.098 | 4 | 4.294 | 195 | 100.852 | .000 |

A high value of F ratio between and within cluster and a low significance value implies that the variables vary significantly across clusters. As can be seen from Table VI, all variables vary across the clusters, with variable 2 varying the least and variable 5 varying the most.

### 3.3. Analysis of Hierarchical vs. K-Means Clustering Results

The comparative analysis if Hierarchical and K-Means showed that the results of a hierarchical cluster analysis for a 5 cluster solution compared with a K-means clustering solution for 5 clusters. The numbers below represent the case membership for 5 hierarchical clusters. Those numbers highlighted in bold represent those members whose cluster has changed in a subsequent K-means analysis. Those which are not highlighted represent those cases which remain in the same cluster in both hierarchical and K-means analysis.

***Membership of Cluster 1:***

1,2,**3,4,5**,**6**,8,**9,11**,14,15,**16**,18,19,**20**,22,**23,24,26**,27,**28,29**,30,33,**36,37**,**38**,39,**41,42**,43,44,**47,48**,**51,52,53**,54,**60,66,67,71**,72,**73,75,76,79**,80,82,85,89,92,**93,97,99**,102,**103,105**,106,108,112,**113**,114,**115,116,121**,122,**123,125**,128,130,131,**132**,133,**136,138**,145,**147,154,155,156,158**,159,**160,161,167**,169,**172,174,177,178,180**,**184,185**,187,**188,192,194**,**196**,197,**199,200**.

Remarks: Total members in hierarchical solution: 102.
Total number of members in K-means solution: 45
Number of common members: 36

***Membership Cluster 2:***

7,10,**12**,13,17,21,32,**34**,35,40,45,46,**49,50**,55,**56**,58,**59**,61,**62**,63,**64**,65,69,**74,78**,**83**,84,**86**,90,**91,95,96**,98,**100**,101,104,107,**111,117,118,119**,120,**124**,**126**,127,129,**134**,135,137,**139,142**,**143,144,146,148**,**150**,152,**153**,162,163,**164**,165,**166,168,170**,171,**173,175**,**176,179,181**,**182**,186,189,**190,191**,195,**198**.

Remarks: Total members in hierarchical solution: 79
Total number of members in K-means solution: 39
Number of common members: 34

***Membership cluster 3:***

**25,31,57,68,70,87,140,151,193**.(all cases have become members of cluster 4 of K-means )

Remarks: Total members in hierarchical solution: 9
Total number of members in K-means solution: 38
Number of common members: 0

***Membership Cluster 4:***

**77,81,94,110,141,183**.

Remarks: Total members in hierarchical solution: 6
Total number of members in K-means solution: 36
Number of common members: 0

***Membership Cluster 5:***

**109,149**. (Both cases have become members of cluster 2 in K-Means)

Remarks: Total members in hierarchical solution: 2
Total number of members in K-means solution: 42
Number of common members: 0

***Overall Comparison between Hierarchical and K-Means:***

Total number of cases in test = 200.
Number of cases falling in common clusters = 45 +39 = 84
Number of cases falling in different clusters = 200-84= 116.

Thus more than 50% of the cases in our test are apportioned into unlike clusters when put through a Hierarchical and K-means analysis subsequently. This suggests that there are significant discrepancies between the results of a K-means and Hierarchical analysis, even when the number of clusters for a K-means analysis is chosen after looking at the results of an initial Hierarchical analysis as explained before. Thus there still exists the need to find ways to bring the results of a Hierarchical analysis closer to a K-means analysis to give more legitimacy to the cluster analysis technique as a whole. One way could be to do a K-means clustering after every step in a Hierarchical cluster analysis. This way it would be possible to transfer cases across clusters if the need arises and the difference between a K-means outcome and a hierarchical outcome ought to be reduced. However this would require a more complex clustering algorithm which is not available for the time being.

## 4. Conclusions

There is no perfect method for setting standards on educational tests. However, the cluster analysis procedure can provide additional information that can be useful for helping set standards. If test data are available, cluster analysis can be used to help select potential borderline, proficient, below proficient, and other groups of examinees that are typically selected using only expert judgment. Thus, the performance of examinees in specific clusters can be compared to those identified using subjective judgment only. Thus, such analyses could be valuable in helping evaluate the results of both test-centered and (other) examinee-centered methods.

Though the clustering approach does not remove subjectivity from the standard-setting process, it does provide subject-matter experts and test developers with a quantitative method for determining different groups of test takers. A potentially desirable feature of the cluster analysis approach is that it provides different options for setting cut-scores. For example, the interval of overlap between examinees in adjacent clusters could be used to select a cut-score interval rather than a specific cut-score. Such an interval provides flexibility to policymakers who must consider politics, resources, and other factors when deciding where to set a cut-score. Similarly, comparing cut-scores resulting from cluster-defined contrasting and borderline groups allows for the evaluation of competing cut-scores. Thus, clustering procedures can provide a set of potential cut-scores, the elements of which can be further evaluated by content experts, psychometricians, and other relevant constituencies who may inform policy decisions.

An attractive feature of the clustering approach is the absence of a unidimensionality requirement. An interesting observation by Sireci (1995) is that by clustering 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 indicates the test is measuring a unidimensional construct. Thus, cluster or factor analysis of examinees rather than of items may provide new insights regarding test dimensionality.

However there are two areas where attention will have to be given for the sake of validity of cluster analysis. These are: a) evaluation of the stability of the cluster solution across samples, and b) external validation of the solutions. These two evaluations are necessary to ensure the cluster solutions are stable and meaningful rather than artifactual. Future applications with larger sample sizes should consider replicating the analyses over several samples. For instance one way could be to use cross tabulation, in which the available data is divided into two sets and a clustering model is evolved that is compatible with the score distributions in the first set and then that very same particular clustering model is applied to the other data set to see if it also fits that nicely.

Future research should also explore other methods for deriving cut-scores from cluster analysis solutions. For example, given a score interval that seems to best separate clusters differing in proficiency, the score within this interval associated with the greatest test information (i.e., lowest conditional standard error of measurement) may be chosen as the cut-score. 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.

## References

1. Angoff, W. H. **Proposals for theoretical and applied development,** Applied Measurement in Education, 1(3), 1988, pp. 215-222
2. Cizek, G. J. **An NCME instructional module on setting passing scores,** Educational Measurement: Issues and Practice, 15(2), 1996, pp. 20-31
3. Cizek, G. J. **Setting performance standards: Concepts, methods and perspectives,** Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers, 2001
4. Jaeger, R. M. **Certification of student competence,** in: Linn, R.L. (ed.) "Educational Measurement (3rd ed.)", New York: Macmillan, 1989, pp. 485-514
5. Jaeger, R. M. **Setting performance standards through two stage judgmental policy capturing,** Applied Measurement in Education, 8, 1995, pp. 15-40
6. Kane, M. **Validating the performance standards associated with passing scores,** Review of Educational Research, 64 (31), 1994
7. Sireci, S. G. **Using cluster analysis to facilitate standard setting,** Applied Measurement in Education, 12(3), 1999, pp. 301-325
8. Sireci, S. G. **Using cluster analysis to solve the problem of standard setting,** Paper presented at the meeting of the "American Psychological Association", New York, 1995
9. Van der Linden, W. J. **Internationalization in educational measurement,** Educational Measurement: Issues and Practice, 13, 4, 1994

a. * * * **Statistical Package for Social Sciences (2010),** SPSS for Windows (Version 18.0) [Computer software].

10. http://www.statsoft.com/textbook/stcluan.html, accessed in December 2010.
11. http://www.psychstat.smsu.edu/MultiBook/mlt04m.html, accessed in January 2011.
12. http://149.170.199.144/multivar/hc.htm, accessed in January 2011.

[1] Whereas the Reckase charts provide information about the probability of an examine with a certain test mark scoring correctly on a certain item in the test, cluster analysis groups alike students.