The Application of Cluster Analysis in Test Validation for Standard-based High-stakes Assessment

Fen-Lan Tseng, Ming-Chiu Chang, Jia-Min Chiou, Fu-Yuan Hsu
Research Center for Psychological and Educational Testing, National Taiwan Normal University, 162, HePing East Road Section 1, Taipei, Taiwan

Abstract - Test validity is a property of the interpretation assigned to test scores. To provide an objective validating evidence for a standard-referenced assessment is especially important. In this study we utilize a statistical technique, cluster analysis, to explore the validity of one of the expert judgement technique- Yes/No Angoff standard setting method. We first segregated each examinee ability cluster using the hierarchical clustering (HC). Assume that each ability cluster is a Gaussian distribution and that the distribution of each test subject data can be modeled by mixture of Gaussians (MoG), where the mean, variance and the proportion of each cluster were initialized by the HC results. Finally, the ability clustering was implemented by the expectation maximization (EM) method. The results from the traditional standard-setting procedure and cluster analysis were compared. The study concludes that cluster analysis appears useful for helping to set standards on educational tests. In addition, it suggested that cluster analysis could be applied as a support tool to provide validating information in the process of standard setting for high-stakes achievement tests.

Keywords: standard setting; hierarchical clustering analysis; validation; EM algorithm

1. Introduction

Setting standards of performance on assessments is a ubiquitous task in education licensure, certification, and credentialing [4]. In an educational setting, it often refers to the process of establishing one or more cut scores of an examination to classify examinees into groups with different learning achievement levels. Standard setting is one of the most important tasks in the test development, administration, and reporting process, especially for high-stake exams. The importance of standard setting is due to the consequences that can be associated with the results of classification. The consequences of high-stake educational testing can trigger student retention, denial of a diploma, or prevent an applicant from attending college of his or her choice. Therefore, how to set appropriate cut scores to reduce the proportions of false-positive and false-negative classification errors is standard-setters’ responsibility and obligation.

One of the ways to classifying standard setting methods was suggested by Jaeger who proposed two categories of methods: test-centered and examinee-centered [9]. Test centered methods involve the use of expert panelists to scrutinize each item of the test and to make judgements regarding the probable levels of performance that borderline proficient test takers will exhibit on the items. The most popular test-centered method is the Angoff and its modifications. In K-12 education contexts, where tests increasingly comprise a mix of multiple-choice and constructed-response format items, a more recent variation of the basic Angoff approach—called the “Extended Angoff method” —has been developed and used in mixed-format assessment [5]. On the other hand, examinee-centered methods use subject-matter experts to evaluate examinees rather than items. The borderline group method is one of them, which uses experts to select a group of test takers who are considered marginally proficient. The median test score for this borderline group is then used as the relevant cut score [2]. Both types of standard-setting methods rely mainly on content expert judgement and have been criticized lack of demonstrated reliability and external validation.

Consequently, validation of a standard-setting method as well as validity study of the method are required. It takes continuous collection of validity evidence from diverse perspectives to qualify test results as credible and to address concerns over potential inaccuracy in man-made judgments. Three kinds of validity evidence were proposed by [14,15]; procedural, internal and external. External evidence is based on comparisons with external sources, e.g. other measurements of the same knowledge and/or skills, results from other standard-setting procedures, and group distribution when the test is given. A cut-score is viewed as and more trustworthy if different standard-setting procedures result in similar performance standards [12].

Although each standard-setting method has its theoretical appeal, there are critical limitations for all. Guidelines to evaluate standard-setting studies have been proposed by researchers [3,14]. It is also suggested that standard-setting studies can be improved by incorporating validity checks on the resulting standard and by using more than one method. The purpose of this study was to extend our previous study [19] to a mixed-format assessment context. We apply a statistical technique, cluster analysis, to explore the validity of the most frequently used test-centered standard setting method—the Angoff method—and the extension of Angoff method for mixed-format tests. Since both the Angoff and its modification are standard setting techniques based on the expert judgement technique, cluster analysis provides an objective alternative to investigate the validity of expert judgment-based methods of setting cutoff test scores [23]. It should be notice that, cluster analysis alone could be used to set standards on a test, if subject matter experts are available to interpret the clustering results, or if defensible external criterion data are available to validate the results [18].

In this study, three ability levels of examinees, namely, Proficient, Basic, and Below Basic, were classified. We first segregated each examinee ability cluster using the Ward’s method [21] (minimum-variance method) in the hierarchical
clustering (HC) on the random sampled data of each test. Next, each ability cluster is assumed to be a Gaussian distribution so that the distribution of each test subject data can be modeled by mixture of Gaussians (MoG) \[24\]. The classification results obtained from HC were used to initialize the mean, variance and the proportion of each cluster followed by the expectation maximization (EM) method to iteratively update the parameters of each ability cluster until convergence.

In the present study, efforts were made to compare the similarity or differences in cut scores obtained by the Angoff procedure and cluster analysis in the context of a national standardized achievement program in Taiwan, the Comprehensive Assessment Program (CAP) for junior high school students. As the prevalence of mixed-format examinations continues to increase, it is likely that standard setting approaches to deal with these types of assessment will receive more attention, especially in high-stake large-scale assessments. With the validation of the statistical cluster solution, empirical evidences can be provided to support the cutoff scores based on the expert judgement methods.

2. Method

2.1 Data

Examinees response data from the Comprehensive Assessment Program (CAP) of year 2015 was employed in the study. CAP is designed to assess junior high school students' learned knowledge and skills in the junior high school curriculum in Taiwan. The test program covers six subjects: Chinese, English (with two subtests of reading and listening), Mathematics, Social Studies, Natural Science, and Writing. A standards-referenced reporting system has been developed for CAP. Examinees' performances are reported with reference to a set of standards. The performance is reported in three levels (levels Proficient, Basic, and Below Basic), with Proficient being the highest. The Comprehensive Assessment Program is administered once each year and is considered high-stake because the test scores are required for admissions to some senior high schools and vocational high schools in Taiwan.

Among the five tests, the Math test was the only mixed-format assessment, with 25 multiple-choice items and two constructed-response items (with four score points each 0-3) administered in the year of 2015. We have done cluster analysis on a random sample of 5000 examinees' response data from Chinese, English, Mathematics, Social Studies, and Natural Science. The analysis was conducted separately for each of the five testing subjects. After the clustering procedure, the score ranges for each ability group were determined. The cut scores (i.e. the number of corrects of the minimally proficient examinees of Basic and Proficient levels should obtain) were analyzed and tabulated.

Among the testing subjects, the English test score was calculated based on both reading and listening subtest scores, with 80% and 20% weighting of the total combination. A compensatory scoring system was applied for the scoring of the English test, i.e. high scores on the reading subtest can compensate for low scores on the listening subtest. Therefore, to obtain the cut score for the English test, cluster analysis was conducted separately for reading and listening tests, and weighted cut scores for the Basic level was calculated based on the following equation, where \( \text{Reading}_{\text{Basic}} \) and \( \text{Listening}_{\text{Basic}} \) indicate the cutoff for the Basic level in the reading and listening test, respectively. The weighted cut score for the Proficiency level is obtained in the same fashion.

\[
\text{English}_{\text{Basic}} = \frac{\text{Reading}_{\text{Basic}} \times 80}{\text{Total reading items}} + \frac{\text{Listening}_{\text{Basic}} \times 20}{\text{Total listening items}}
\]

The Math test is the only mixed-format assessment, the standard setters applied Yes/No Angoff method for multiple-choice items and the extended Angoff method \[10,11\] for the constructed-response (CR) items to set the cut scores. The Math test score was calculated based on composite scores from multiple-choice items and constructed-response items, with 85% and 15% weighting respectively. Since there are only two constructed-response items, cluster analysis was conducted with the weighted combination scores which were calculated based on the sum of the total correct multiple-choice items multiplied by 3.4 (85/25) and the total scores of constructed-response items multiplied by 2.5 (15/6). The weighted cut scores for the Basic level was calculated based on the following equation, where (Multiple choices) and (scores in CR items) indicate the cutoffs for the Basic level in multiple-choice items and the constructed-response items, respectively. The weighted cut score for the Proficiency level is obtained in the same fashion.

\[
\text{Math}_{\text{Basic}} = \frac{(\text{Multiple choices})_{\text{Basic}} \times 85}{\text{Total items in multiple-choice}} + \frac{(\text{Scores in CR items})_{\text{Basic}} \times 15}{\text{Total scores in CR items}}
\]

2.2 Extended Angoff Method

To set the cut scores for the math CR questions, an extended Angoff was applied. When the subject experts estimate the probability that borderline test takers will answer a multiple-choice correctly in traditional Angoff method, they are actually estimating the expected average score of the borderline test takers for the question. Therefore, a straightforward extension of the Angoff’s method for CR items is to have the experts estimate the average score that borderline test takers will answer. Each of the two CR questions of the Math test of CAP is scored polytomously on a 0-3 rubric with the total possible points ranging from 0 to 6. During the three iterations or rounds of the item judgement process, some math content experts may judge that the minimally competent examinee of Proficient level would obtain 1 out of 3 points on average; other judges may estimate the average score of the borderline examinee of Proficient level to be 2 out of 3. The final
estimated scores of each round for the borderline examinee of **Proficient** level are obtained by averaging the estimated scores from all judges. The extended Angoff method can be combined with the traditional Angoff or yes/no Angoff when the test is composed of both multiple-choice and constructed-response questions [13].

### 2.3 Yes/No Angoff method

In the basic form of the Angoff, the judges are instructed to imagine a minimally competent examinee and to estimate the probability that this borderline examinee will answer each test item correctly [5,6]. Typically, the judges are instructed to think of a pool of 100 candidates who would just barely meet the performance criteria. Working independently, the panel judges then make an estimation on what proportion of that sample of minimally acceptable candidates would answer each item in the test correctly. These estimated p-values are summed and usually denoted as the Minimum Passing Level for judge (MPLs). The MPL symbolizes an individual judge’s cut score for the test. The mean of all MPLs or cut scores is the final cut score for the test. A low standard error of the MPLs is desirable since it indicates better consistency among the panel judges.

The main concept of the Angoff method is to estimate the proportion of that minimally competent examinees would correctly answer a test item. To simplify the procedure, Impara and Plake [13] proposed the standard-setters decide whether a single borderline examinee would or would not answer an item correctly. Known as the Yes/No Angoff method or procedure, the rationales for the modified version are that it is much easier for a standard-setter to think of a single examinee than a pool of them and to make a simple yes/no decision. Setting cut scores repeatedly is regarded as a desirable characteristic of a judgmental process [8]. The time between rounds is arranged for the standard setting panel discussion. The purpose of the panel discussion is to enhance the agreement among the panelists. Use of two or three iterations have been reported in the literature.

It remains a controversial issue over whether it is appropriate or not to provide normative data to the standard setting judges during the Angoff procedure. Presentation of quantitative data based on the examinee performance has been shown to improve the inter-rater reliability. The standard setting panelists are usually provided with normative data prior to the final iteration of estimating item p-values. Busch and Jaeger [1] found that presentation of examination results to members to discuss on the disagreement, especially between the first and second round.

The Comprehensive Assessment Program (CAP) of year 2015 was administered in the month of May and the score reports were provided to each examinee in June 5th, 2015. For each of testing subject, except Chinese Writing, a standard setting meeting with the application yes/no Angoff method and extended Angoff for the constructed-response items of the Math test was held in the end of May, 2015. The coefficient alpha for each test is Chinese .91, English listening .88 (with 21 items), English reading .94, Math 0.88, Social Studies .93, Science .94, respectively.

### 2.4 Hierarchical Clustering

Cluster analysis can be used to group test takers into homogeneous clusters with respect to the proficiency measured. Previous researches have applied the grouping results based on k-means clustering as the initial cluster centroid [16,18,20]. This study applied expectation maximization (EM) algorithm instead to implement the ability clustering procedure. Compared with k-means clustering, EM algorithm is less biased toward spherical clusters of about equal size [9] and is more preferable in validating the clustering of the examinees of the test subjects of CAP which show highly unequal cluster size based on the Angoff standard setting method.

Hierarchical clustering is one of methods way to segregate a set of data into different clustering. The computational procedure of HC includes the following three steps. The first step is to calculate the dissimilarity between every pair of samples in the data set. The second step is to group the samples into a binary and hierarchical cluster tree. The third step is to determine the level of clustering by cutting the hierarchical tree. Various algorithms have been implemented to calculate the dissimilarity and create the cluster tree [23]. In this study, the dissimilarity was calculated using the Euclidean distance defined by

\[
d_{ij} = (x_i - x_j)^2
\]

where \(x_i\) and \(x_j\) were two different examinees. Let us define the numbers of examinees in cluster \(p\) and \(q\) by \(n_p\) and \(n_q\), respectively, the centroids of clusters \(p\) and \(q\) by \(\bar{x}_p\) and \(\bar{x}_q\), respectively. Specifically,

\[
\bar{x}_p = \frac{1}{n_p} \sum_{i=1}^{n_p} x_{pi} \quad \text{and} \quad \bar{x}_q = \frac{1}{n_q} \sum_{j=1}^{n_q} x_{qj}
\]

where \(x_{pi}\) is the \(i\)th examinee in cluster \(p\) and \(x_{qj}\) is the \(j\)th examinee in cluster \(q\).

In order to construct the cluster tree, we utilized the Ward’s method to compute the distance between two clusters. Tinsley and Brown [21] pointed out that the Ward’s method is one of the most robust and accurate methods in hierarchical clustering analysis. It is emphasized here that the clustering results arrived at through the Ward's method simply serve as the initial values of estimated parameters in the EM algorithm,
in order to optimize the clustering results while assuming each ability cluster is a Gaussian distribution. The Ward’s method was defined by

\[
d^2(p, q) = n_p n_q \frac{(\bar{x}_p - \bar{x}_q)^2}{(n_p + n_q)}
\]

where \(d^2(p, q)\) is the distance between two clusters. In this study, calculations were made using the "ward" option in the linkage.m function in Matlab 2017a.

The Expectation Maximization (EM) Algorithm is a method widely used in Data Clustering. In statistical calculations, this method is applied to find the maximum a posteriori probability for parameters in a probability model, and the parameter estimate in the probability model depends on the non-observable latent variables. It is the assumption of this study that each cluster is a Gaussian distribution, and that the probability of correct items in a subject by any given test taker is a mixture of Gaussian distributions. The objective of this study is to uncover the ability cluster of each examinee, i.e. latent variable. In the Expectation Maximization Algorithm, there are two alternate steps: E-step and M-Step. E-step, the first step, uses the estimated Gaussian Mixture Model parameters to re-calculate the probability of the number of correct items by each test taker being classified into different performance levels, which we call the posteriori probability of performance level. M-step, the second step, uses the posteriori probability of performance level to re-optimize the Gaussian Mixture Model parameters. The Gaussian Mixture Model parameters determined in M-step are used to calculate the posteriori probability for the next E-step, and the two-step procedure is performed alternately and iteratively. It should be highlighted that the results of HC analysis can be used to estimate the mean and variance of each cluster, as well as to initialize the Gaussian Mixture Model parameters in the EM method, after taking into account the proportion of examinees in certain cluster among all examinees. In this study, the HC was initiated with 3 clusters. The HC result was used to calculate the mean, variance, and the proportion of each class which were initial parameters in the model of MoG.

2.5 Expectation Maximization Algorithm

This study applies EM to look for unknown parameters in the Gaussian Mixture Model. Assuming each ability cluster for certain subject demonstrates Gaussian distributions, and the number of correct items is \(x_n, n = 1, \ldots, N\), and \(x_n\) is a mixture of Gaussian distributions, the posteriori probability distribution of the sample is as follows:

\[
p(x_n|\mu_i, \sigma_i, \omega_i) = \sum_{i=1}^{K} \omega_i \frac{1}{\sqrt{2\pi\sigma_i}} \exp \left[ -\frac{(x_n - \mu_i)^2}{2\sigma_i^2} \right]
\]

(6)

In the formula, \(i = 1, \ldots, K\) denotes the labels for different ability levels, \(1 / \sqrt{2\pi\sigma_i} \exp \left[ -\frac{(x_n - \mu_i)^2}{2\sigma_i^2} \right]\) is Gaussian distributions, which can be represented as \(g_{x_n}(\mu_i, \sigma_i)\) for the purpose of simplification, where \(\mu_i\) denotes the mean, \(\sigma_i\) denotes the variance, \(\omega_i\) denotes the proportion of the \(i\) cluster in the Gaussian Mixture Model, and the total of all the \(\omega_i\) is 1, i.e. \(\sum_{i=1}^{K} \omega_i = 1\). Clustering analysis, conducted through the EM algorithm, is used to estimate the mean and variance in each ability cluster, and the proportion of distribution, so that the probability model of the mixture of Gaussian distributions can fit the response data of examinees. In this study, all ability clusters demonstrate Gaussian distributions. As there are 3 clusters, a total of 9 parameters have to be estimated: \(\{\mu_1, \sigma_1, \omega_1\}, \{\mu_2, \sigma_2, \omega_2\}\) and \(\{\mu_3, \sigma_3, \omega_3\}\). The initial values of the 9 parameters can be arrived at by calculating the results of the 3 clusters through HC analysis. With the 5 subjects in CAP, which is the target of this study, each subject has gone through independent HC analysis along with EM analysis. The EM algorithm consists of 2 alternate and iterative steps: E-step and M-step. Let the number of correct items of each examinee for a specific subject denoted by \(x_n, n = 1, \ldots, N\). We implemented the data segregation using EM algorithm by fitting the data distribution into a mixture of Gaussian models. There are two alternating iterative steps in the EM method, namely, E-step and M-step [24]. In the E-step, we compute the posteriori probabilities of ability classes \(p(i|x_n, \theta^{t-1})\) derived from previous \((j-1)\)th iteration

\[
p(i|x_n, \theta^{j-1}) = \frac{\pi^{j-1}_i g_{x_n}(\mu^{j-1}_i, \Sigma^{j-1}_i)}{\sum_{i=1}^{K} \pi^{j-1}_i g_{x_n}(\mu^{j-1}_i, \Sigma^{j-1}_i)}
\]

(7)

where \(i = 1, \ldots, K\) represent labels of different proficient levels, \(g_{x_n}(\mu_i, \Sigma_i)\) represents the Gaussian density with the mean \(\mu_i\) and variance \(\Sigma_i\) of ability level \(i\), \(\pi_i\) denotes the proportion of each cluster \(i\), \(\sum_{i=1}^{K} \pi_i = 1\), and \(\theta^{j-1}\) denotes the parameters \(\{\mu^{j-1}_i, \Sigma^{j-1}_i, \pi^{j-1}_i\}\) at the \((j-1)\)th iteration. In the M-step, we estimate the following parameters \(\theta^j\) in which

\[
\mu^j_i = \frac{\sum_{n=1}^{N} p(i|x_n, \theta^{j-1}) x_n}{\sum_{n=1}^{N} p(i|x_n, \theta^{j-1})},
\]

\[
\Sigma^j_i = \frac{\sum_{n=1}^{N} p(i|x_n, \theta^{j-1}) (x_n - \mu^j_i)^y}{\sum_{n=1}^{N} p(i|x_n, \theta^{j-1})},
\]

\[
\pi^j_i = \frac{1}{N} \sum_{n=1}^{N} p(i|x_n, \theta^{j-1})
\]

(8)

The number of classes \(K=3\) is predetermined by educational policy in Taiwan. The ability level of each examinee was determined by the maximum of \(p(i|x_n, \theta^{j-1})\).

3. Results

1.1 Comparisons of Two Methods

Table. 1 demonstrated the cutoff scores by the Angoff method and cluster analysis, respectively. It reveals a trend
that the objective statistical technique tends to yield lower cutoff scores for Proficient achievement levels across tests, except for Math test and higher cutoff scores for Below Basic achievement levels across tests, except for Math test. The two methods yielded the most similar cutoff scores for Math test compared with the rest. Both methods yielded the same cutoff score, i.e. a weighted score of 76.2, for the Proficient achievement level in Math test.

Table. 1 Comparison of Cutoff Scores by The Angoff method and Cluster Analysis (MoG/EM)

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th>English</th>
<th>Math</th>
<th>Social Studies</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Angoff</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total items/</td>
<td>48</td>
<td>100</td>
<td>100</td>
<td>63</td>
<td>54</td>
</tr>
<tr>
<td>Total weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scores</td>
<td>48</td>
<td>100</td>
<td>100</td>
<td>63</td>
<td>54</td>
</tr>
<tr>
<td>Proficient</td>
<td>41</td>
<td>88</td>
<td>76.2</td>
<td>53</td>
<td>47</td>
</tr>
<tr>
<td>Cutoff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>20</td>
<td>40.38</td>
<td>36.3</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td><strong>MoG/EM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total items/</td>
<td>48</td>
<td>100</td>
<td>100</td>
<td>63</td>
<td>54</td>
</tr>
<tr>
<td>Total weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scores</td>
<td>48</td>
<td>100</td>
<td>100</td>
<td>63</td>
<td>54</td>
</tr>
<tr>
<td>Proficient</td>
<td>40</td>
<td>86</td>
<td>76.2</td>
<td>51</td>
<td>46</td>
</tr>
<tr>
<td>Cutoff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>26</td>
<td>52.19</td>
<td>33.1</td>
<td>34</td>
<td>27</td>
</tr>
</tbody>
</table>

Then, examinees of 2015 CAP were classified into three performance levels, Proficient, Basic and Below Basic, based on their numbers of correctly answered items (Chinese, Social Studies, and Science) or weighted scores (English and Math) in all subjects using cutoff scores in Table. 1. Table. 2 showed the percentages of examinees of different achievement levels classified by the Angoff method and cluster analysis, respectively. Take the Chinese subject for example, under the Angoff method, examinees at Proficient level accounted for 18.06% of all examinees, those who were at Basic level accounted for 63.96%, and those at Below Basic level accounted for 17.98%. In comparison with cluster analysis, the percentages of examinees at the three levels accounted for 21.23%, 44.13% and 34.64% of all examinees respectively. Table. 2 reveals that cluster analysis resulted in much smaller percentages of examinees for Basic levels across tests, except for the Math test. The objective statistical technique yielded more percentages of examinees for Below Basic levels across tests, except for Math test.

Angoff method tends to be stricter at the highest achievement level, and therefore some examinees classified as Proficient level with cluster analysis were considered Basic level under the Angoff method. The 3.17% in Table. 2 for the Chinese subject is an example of this shift. On the other hand, Angoff method is more lenient at the lowest achievement level, so some examinees classified as Basic level were considered Below Basic level with cluster analysis. The 16.66% in Table. 2 for the Chinese subject is an example of this shift. The cutoff scores for Proficient level in Math are the same under the two methods, so both percentages were 15.08%. Here again Angoff tends to be stricter but at the Basic level cutoff this time, so 3.63% of examinees who were Below Basic level under Angoff were classified as Basic level with cluster analysis.

Table. 2 Comparison of Percentages of Examinees of Different Achievement Levels Classified by the Angoff method and Cluster Analysis (MoG/EM)

<table>
<thead>
<tr>
<th></th>
<th>Proficient</th>
<th>Basic</th>
<th>Below Basic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chinese</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient</td>
<td>18.06</td>
<td>3.17</td>
<td>0.00</td>
<td>21.23</td>
</tr>
<tr>
<td>Basic</td>
<td>0.00</td>
<td>44.13</td>
<td>0.00</td>
<td>44.13</td>
</tr>
<tr>
<td>Below Basic</td>
<td>0.00</td>
<td>16.66</td>
<td>17.98</td>
<td>34.64</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>18.06</td>
<td>63.96</td>
<td>17.98</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>English</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient</td>
<td>18.15</td>
<td>2.76</td>
<td>0.00</td>
<td>20.91</td>
</tr>
<tr>
<td>Basic</td>
<td>0.00</td>
<td>34.17</td>
<td>0.00</td>
<td>34.17</td>
</tr>
<tr>
<td>Below Basic</td>
<td>0.00</td>
<td>11.72</td>
<td>33.20</td>
<td>44.92</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>18.15</td>
<td>48.64</td>
<td>33.20</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Math</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient</td>
<td>15.08</td>
<td>0.00</td>
<td>0.00</td>
<td>15.08</td>
</tr>
<tr>
<td>Basic</td>
<td>0.00</td>
<td>51.70</td>
<td>3.63</td>
<td>55.33</td>
</tr>
<tr>
<td>Below Basic</td>
<td>0.00</td>
<td>0.00</td>
<td>29.60</td>
<td>29.60</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>15.08</td>
<td>51.70</td>
<td>33.22</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Social</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient</td>
<td>16.69</td>
<td>4.48</td>
<td>0.00</td>
<td>21.16</td>
</tr>
<tr>
<td>Basic</td>
<td>0.00</td>
<td>39.90</td>
<td>0.00</td>
<td>39.90</td>
</tr>
<tr>
<td>Below Basic</td>
<td>0.00</td>
<td>24.18</td>
<td>14.76</td>
<td>38.94</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>16.69</td>
<td>68.55</td>
<td>14.76</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proficient</td>
<td>13.90</td>
<td>1.65</td>
<td>0.00</td>
<td>15.54</td>
</tr>
<tr>
<td>Basic</td>
<td>0.00</td>
<td>33.83</td>
<td>0.00</td>
<td>33.83</td>
</tr>
<tr>
<td>Below Basic</td>
<td>0.00</td>
<td>27.82</td>
<td>22.81</td>
<td>50.63</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>13.90</td>
<td>63.30</td>
<td>22.81</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Table. 3 demonstrated classification consistency between Angoff and cluster analysis. The percentage agreement between the two methods ranged from 70.53% to 96.37%. Classification results for all the subjects, except Social Studies and Science, showed high consistency at 80% and above, while the consistency in Math was as high as around 96%. The high consistency between the two methods was also shown with Kappa coefficients ranging from 0.55 to 0.94. Based on Fleiss's (1981) practice in determining consistency using Kappa coefficients, the consistency is almost perfect when the coefficient goes beyond 0.81, substantial if between 0.61 and 0.81, and moderate if between 0.41 and 0.60. Consequently, the classification consistency was almost perfect in Math, substantial in Chinese and English, and moderate in Social Studies and Science. As indicated in Table. 3, high classification consistency was shown between the two
methods, excluding Social Studies and Science, with percentage agreement at 80% and above, Kappa coefficient at 0.6 and above. Therefore, there is a high level of classification consistency between the two standard setting methods adopted in this study.

Table 3 Classification consistency by The Angoff method and Cluster Analysis (MoG/EM)

<table>
<thead>
<tr>
<th>Subject</th>
<th>Percentage Agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>80.17%</td>
<td>.68***</td>
</tr>
<tr>
<td>English</td>
<td>85.52%</td>
<td>.78***</td>
</tr>
<tr>
<td>Math</td>
<td>96.37%</td>
<td>.94***</td>
</tr>
<tr>
<td>Social Studies</td>
<td>71.35%</td>
<td>.55***</td>
</tr>
<tr>
<td>Science</td>
<td>70.53%</td>
<td>.55***</td>
</tr>
</tbody>
</table>

***p<.001

3.2 Mixed-format Assessment

Among the five tests in this study, the Math test is the only mixed-format assessment, with 25 multiple-choice items and two constructed-response items (with four score points each). The standard setters applied Yes/No Angoff method for multiple-choice items and the extended Angoff method for the constructed-response items to set the cut scores. Table 3 showed the cluster analysis provides a desirable alternative for the validity checks on the results obtained based on the traditional standard setting methods, especially for the Math test. With mixed test formats, the two methods yielded higher classification in the 2015 Math test than in the 2014 Math test [19]. The results of the two methods are apparently more consistent in Math (mixed test format) than in other subjects in 2015, and also more consistent in the 2014 Math test (27 dichotomous items), suggesting cluster analysis may be more suited for mixed-format tests.

4 Discussion

In this study we investigated the validity issue in setting standards for the CAP which is considered high-stakes in terms of its applications in admission-decision making and the accountability involved. Since the standard-referenced scoring for a mixed-format high-stakes large scale assessment has never been carried out before in Taiwan, efforts should be made to establish more empirical evidences for the sake of validity.

4.1 Different Patterns of Ability Grouping Across Tests

Table 3 indicated that the classification consistency in Social Studies and Science tests was obviously lower than in the other three subjects. Social Studies and Science happen to both have sub-subjects: history, geography and civic studies in Social Studies; physics, chemistry, biology and earth science in Science. For Social Studies and Science tests, the Angoff method was applied to the sub-subjects first, and the results from each sub-subject were consolidated to form the final cutoff score for the subject, whereas cluster analysis was applied directly to the two subjects. As cluster analysis is used to determine the cluster of an examinee by calculating the distance among examinees, it is worth further studies to investigate whether the two methods’ inconsistency in Social Studies and Science can be attributed to the fact that both subjects have sub-subjects and that the inconsistency arises from performance variance among the sub-subjects from the same examinee.

For the five tests in the year of 2015, the Angoff method resulted in different percentages of correct items/or weighted scores required to reach the Proficient levels across tests, from 76.2% to 88%. On the other hand, the percentages of correct items/or weighted scores required to reach the Basic levels are also similar across tests, from 35% to 41%. In the mean times, solutions based on of cluster analysis showed the percentages of correct items/or weighted scores required to reach the Proficient levels are also different across tests, ranging from 76% to 86%. As for the percentages of correct items/or weighted scores required to reach the Basic levels, the cutoff range for standards are also similar across tests, from 50% to 54%, except the Math test. The statistical clustering method apparently yield higher cutoff for Basic levels and lower for the Proficient, except for the Math test.

A key in this study is the assumption that each ability cluster shows Gaussian distributions. As past literature has confirmed that the Ward HC method can achieve good clustering results, such results are simply used in this study to estimate initial parameters of each Gaussian distributions model. The clustering of the parameters, comprising mean, variance, and proportion of examinees in each cluster, is then optimized through EM algorithm. The initial values can also be estimated with the clustering results obtained through non-HC clustering (such as k-means, [17]). Although statisticians have mixed views toward clustering analysis, what we want to emphasize is any clustering analysis capable of good clustering results can be performed in similar practical applications.

The focal point of this study is not the attempt to develop an academically original method of classification but the application of existing and mature classification method to verify the validity of a standard-setting approach. Adopted in this study is the EM algorithm to optimize the estimated parameters in the Gaussian mixture model, so that the model can fit examinees’ response data. The clustering results generated by the HC method are used to calculate initial mean, variance and the proportion of distributions in the Gaussian mixture model used by each Gaussian model. In comparison, Khalid [16], Sireci, Robin and Patelis [18], and Violato, Marini and Lee [22] used HC analysis results as the central point of different clusters under the k-means method before iterative clustering is performed and applied in standard-setting research. Both classification methods utilize the HC analysis to initialize the next step, but the EM algorithm adopted in this study does not lead to the issue of equal cluster size [17] and appears to be more fitting for verifying the actual proportions of Proficient, Basic, and Below Basic in CAP (current expert
judgments designate more than half of the examinees as at the basic level).

The statistical method also showed an extraordinary phenomenon of high percentages of examinees being clustered into the Below Basic levels across tests, except for the Math test. This outcome deserved for future research into whether curriculum of each subject areas set out different standards to achieve.

4.2 Similar Clustering Results for Math Test

In this study, both the expert judgement and cluster analysis methods yield the same cutoff score for Proficient level for Math test. As to the cutoff scores for Basic level, the two methods also provides similar weighted cutoff scores, i.e. 36.3 and 33.3. The reason for the Math test to stand out in the standard setting process is probably attributed to the clarity of the Performance Level Descriptor or Description (PLD) of Math. With a mixed-format design, the classification agreement between the two approaches is the highest among all tests. In light of the objective clusterizing outcomes, standard setters should evaluate the possible causes and provide more convincing evidences for the stakeholders to rely on the scoring outcomes and its following applications.

As cluster analysis is used to determine the cluster of an examinee by calculating the performance gap among examinees, it is worth further studies to investigate whether the two methods' inconsistency in social studies and science can be attributed to the fact that both subjects have sub-subjects and that the inconsistency arises from performance variance among the sub-subjects from the same examinee.

Acknowledgment

This work was supported in part by Ministry of Science and Technology, Chinese Taipei (MOST 105-2511-S-003-033) and was also financially supported by the “Institute for Research Excellence in Learning Sciences” of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

5 References


