How Good is Good Enough?: A Comparison of Three Methods for Establishing Cut Scores on Placement Tests

Sang-Keun Shin
(Ewha Womans University)


The purpose of this study was to determine whether three standard-setting methods would produce sufficiently consistent results for placement decisions. The cut scores were derived from modified Angoff method, borderline group method, and cluster analysis. The results indicate that the cut scores derived from the three standard setting methods did not entirely agree with each other in assigning students into different levels, suggesting that the choice of standard setting method influenced the resulting cut scores. Specifically, the borderline-group method tended to produce lower cut scores than the other two methods. The cluster analysis yielded two cut scores, which were very similar to the cut scores derived from the modified Angoff method at the corresponding levels. The implications of the findings are discussed and avenues for further study are suggested.

**Key words:** placement test, cut score, standard setting

1. INTRODUCTION

One of the primary uses of language tests is to make various decisions such as placement and certification decisions which involve the classification of test takers into ordered performance categories such as *pass* and *fail*, or *advanced* and *intermediate*. When language tests are used for classification purposes, test standards are essential. For example, most university English for Academic Purposes programs place incoming non-native English speaking students into different levels of instruction based on their performance on placement tests. In order to make such placement decisions, English language programs have to first define the level of performance at each placement level and then decide where and how to set cut-off points. But how good is good enough for each placement level, how
should these levels be established, and who should make these decisions?

When standards are set on language tests, we must be concerned with their quality because important consequences are typically tied to them. In educational testing, classification decisions can influence students’ educational opportunities. In licensure/certification testing, the financial future of test takers can be at stake. For these reasons, it is very important that the categorization is a valid one.

The consequences associated with test standards have motivated intensive research into standard setting methods, and a large number of standard setting methods have been developed in the educational measurement literature (see Berk, 1986; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006, for a review). Despite the profusion of standard setting methods, the literature on standard setting concurs that it is difficult to set a defensible standard on a test. Since the score scales for language tests are typically constructed to order examinees along a continuum, there is no simple and obvious way to choose a cut score that supposedly divides students into two distinct categories such as pass and fail. For instance, a TOEFL score of at least 87 on the internet-based test is the minimum required for admission to graduate programs at University of California at Los Angeles (University of California at Los Angeles, 2014). But why 87? Why not 97? Why could it not be set a little bit higher or a little lower? Are the students who scored 86 really different from those who obtained 87 in terms of their academic English language ability?

Since validity pertains to the appropriateness of inferences and decisions that we make on the basis of test scores (Messick, 1989), the validity of a test that is used for classification decisions is also linked to the degree to which cut scores are accurate. Even if a language test is valid for making an inference about test takers’ proficiency for a given purpose, cut scores, which are either too high or too low, will misclassify test takers, and this classification error will result in inaccurate inferences about test takers’ language ability. Since the adoption of the cut scores to assign examinees to performance levels adds a new layer to the interpretation of test scores, setting test standards is a validity issue, and it is thus critical that test developers and users gather evidence to support the appropriateness of the test standards for their given purposes.

Even though cut scores are widely used for making decisions based on language test results, how they were arrived at is seldom documented, and how they were established often remains unquestioned. Since cut scores are an answer to the question of “how good is good enough”, setting cut scores is a measurement issue, and the role of measurement is to provide decision makers with accurate and relevant information so that they can make those decisions as accurately and dependably as possible. As Brown (1996) points out, even though cut scores are often the basis for making important decisions that may strongly affect test takers’ lives, they are not commonly discussed in the language teaching and testing field. As a result, we know little about how to translate performance standards
into defensible and valid cut scores. Cut scores are often set through an overly intuitive process without sufficient theoretical and empirical bases. The question of “how much” is often answered not in terms of a predefined level of performance but in terms of the relative standings of students or even logistical needs. Again, setting cut scores is a validity issue because cut scores that are too high or too low will misclassify students and result in inaccurate inferences about test takers’ proficiency.

Since there are a range of choices among standard setting methods and no agreement on a best method, many studies have compared the results of different standard setting methods. One consistent finding from these studies is that different standard setting methods often produce different standards (see Hambleton & Pitoniak, 2006; Jaeger, 1989, for a review), and this lack of comparability among methods calls for the examination of the validity of methods for setting standards. If cut scores depend heavily on the methods used to derive them, many students could be misclassified on the basis of methodological differences in standard setting procedures that have nothing to do with their abilities.

Language test users must often make decisions about students, particularly for such purposes as admissions and placement. Standard-setting has recently become an important issue in secondary schools in Korea because teachers are required to set standards on achievement tests due to the adoption of achievement standard assessment system. However, to date, with rare exception (Bechger, Kuiper, & Maris, 2009; Griffin, 2001; Hsieh, 2013; Kozaki, 2004; O’Neill, Buckendahl, Plake, & Taylor, 2007; Stansfield & Hewitt, 2005; Xi, 2007), little attention has been paid to standard setting issues in language tests. Alderson, Clapham and Wall (1995) conducted a survey of British examining boards and reported that even though several testing boards appear to have fixed pass marks, no further information was available on how these marks were arrived at. In addition, no justification was offered for the pass marks reported, nor was there any discussion of how variations in the difficulty of the examination from year to year or the quality of the candidature were taken into account when setting and applying these pass marks.

Given the dearth of studies examining the validity of language test performance standards, further investigations are needed in order to explore how to set and validate standards for language tests. This is because “shrouding the decision-making process in mystery or secrecy is not an adequate response to the problem” (Alderson et al., 1995, p. 166). Language testers and test users have a duty to inform test takers and other stakeholders about the procedures they follow to ensure that reasonable decisions have been made, and also to give people confidence in the standards themselves and their meaningfulness for the purpose for which they will be applied. This study attempted to fill the gap in the literature by investigating standard setting procedures for an institutional placement test and to evaluate the appropriateness of the resulting standards by examining the comparability of cut scores derived from three standard setting methods: modified
Angoff method, borderline group method, and cluster analysis.

2. LITERATURE REVIEW

Since cut scores are called standards by many authors, standard setting methods in the measurement literature often refer only to methods for establishing cut scores on tests. A large number of standard setting methods have been documented in the measurement literature, and these methods can be divided into two categories: test-centered methods and examinee-centered methods (Jaeger, 1989; Zieky, 2001). Test-centered methods involve the use of expert judges to examine test items and to make judgments regarding the probable levels of performance that marginally proficient test takers will exhibit on the items. On the other hand, examinee-centered standard setting methods use subject-matter experts to evaluate the work of examinees rather than items. A brief discussion of the most widely used test-centered and examinee-centered methods is provided below. Two statistical approaches which have been applied to establish cut scores on examinations are also briefly discussed.

2.1. Test-Centered Methods

In the test-centered methods, judges set the standard by reviewing the items in the test and deciding on the level of performance on these items that will be considered just adequate to meet each performance standard. Among a number of ways for raters to do this (Angoff, 1971; Ebel, 1972; Jaeger, 1982; Nedelsky, 1954), the most popular test-centered method is the Angoff method (Angoff, 1971) and its modifications (Cross, Impara, Frary, & Jaeger, 1984; Jaeger & Busch, 1984). For items scored dichotomously, the judges in an Angoff study estimate the probability that a borderline examinee (or just passing examinee) will answer an item correctly. For items that are scored polylomously, the judges estimate the expected score of a borderline candidate on the item (Hambleton & Plake, 1995), and cut scores are set on the test score by summing the expected item scores for each judge and then averaging these sums across judges. For example, raters may agree that the probability of a borderline candidate getting Item 1 correct is 50%, whereas for Item 2 it is 75% and for Item 3, 25%. The sum would be divided by 3, and the cut score for this three-item task would thus be 1.5.

Angoff method has been modified in a number of ways to suit individual applications. For example, an iterative process is often used in which judges are given the opportunity to discuss their initial ratings and then modify their ratings (Jaeger, 1989). Often, item statistics or information about the passing and failing rates are presented to judges. In
another modification, judges are asked to choose the percentage of minimally knowledgeable examinees who would know the answer to each item from among seven choices (5, 20, 40, 60, 75, 90, and 95) or to choose ‘Do not know’ (Educational Testing Service, 1976).

Although widely used, there have been recent criticisms that “the Angoff method is fundamentally flawed” (Shepard, Glaser, Linn, & Bohrnstedt, 1993, p. 3). The success of the Angoff procedures depends on how well a group of panelists can conceptualize the ability of a hypothetical borderline candidate and estimate the probability that this borderline test taker will correctly answer each item on a test. After reviewing the standard setting study of the National Assessment of Educational Progress (NAEP), Shepard et al. concluded that it was difficult for judges to clearly envision the knowledge and skills characterizing borderline test takers and to map these levels of knowledge and skills to those required for success on numerous test items. However, the Angoff method has been widely recommended because of the general reasonableness of the standard set, the ease of use, and the psychometric properties of the standard. For instance, Kane (1995) defends the Angoff method as having been used on a host of licensure and certification tests, as well as on numerous state testing programs without major complaints from the participants involved.

2.2. Examinee-Centered Methods

Since examinee-centered standard setting methods require panelists to look at the work of students (i.e., samples of student work), raters typically categorize examinees based on some external criterion assessment of the examinee’s level of performance. The cut scores are then set by identifying points on the score scale that would be most consistent with these categorization decisions. For example, the contrasting-groups method (Zieky & Livingston, 1977), the most popular examinee-centered method (Mehrens, 1994, p. 8), uses experts to select two groups of test takers, one group considered to be above the relevant standard scale and the other group considered to be below this standard. The cut score is then identified as a point somewhere between the distributions of the two identified groups. In general, the test score that results in the fewest false-positive and false-negative classifications is selected as the passing score.

Similarly, in the borderline-group method, judges identify borderline test takers whose performance is right around the performance standard under consideration (i.e., between two performance standards). Some index of central tendency such as the median test score for borderline examinees can then be used as the cut score. Generally, the borderline-group method is working well if most of the borderline-groups’ test scores are clustered close together. On the other hand, if the scores of the borderline-group are spread widely over
the range of possible scores, then the method is not working well because the borderline-
group may include test-takers who do not belong in it.

One advantage of examinee-centered methods is that the cut score is usually based on
the observations of experts who are familiar with test takers. For example, if the standard
being set is for achievement in the class they are teaching, classroom teachers are an
excellent source of information about students’ performance. Another advantage of
examinee-centered methods is that they ask raters to judge the performance of individuals
rather than to perform the less familiar task of rating test items. However, examinee-
centered methods can readily be employed only when experts have experience with the
students. The problem is that such experience may not always be available, and if this is
the case, a special criterion measure needs to be used to classify a large number of
examinees as borderline or as being in one of two contrasting groups. Since there is no
direct way to determine borderline proficiency, the same types of false-positive and false-
negative classification errors associated with standard setting apply to the assignment of
test takers to the borderline-group.

2.3. Statistical Methods

The statistical standard setting methods are different from both test- and examinee-
centered methods in that they involve statistical analysis of test score data rather than
statistical analysis of raters’ judgments. In the educational and psychological testing
literature, cluster analysis and latent class analysis have recently been used to set standards
or assist other standard setting methods.

In one of the most well-known studies, Sireci, Robin, and Patelis (1999) cluster-
analyzed a statewide mathematics test to discover groups useful for envisioning marginally
competent performance as required in test-centered standard setting methods or defining
borderline or contrasting groups used in examinee-centered methods. The results of their
analysis showed that a three-cluster solution reflected the three proficiency groupings the
test was designed to measure. Students within each cluster were also statistically
significantly different from students in other clusters with respect to final math grades. The
researchers also reported that the concordance rate of students who were classified into the
same categories using the cluster analysis and the Angoff procedure was 93%, and the
Cohen’s (1960) kappa (proportion of exact agreement corrected for chance) was .90.

It is important to note that statistical methods do not solve the standard setting problem
because the standard setting problem cannot be reduced to a statistical test (Sireci, 2001).
As applied to standard setting, these methods also have a few limitations. For example,
they will cluster the data regardless of whether truly different groups of examinees are
present, and thus there is a need for external validation data to ensure that resulting clusters
(or latent classes) are qualitatively different from one another. Another problem is that no standards can be set higher or lower than the test takers actually perform because these methods focus on analysis of test response data. Put differently, both methods derive standards based on “what specific groups of test takers have done, rather than according to what they should have done” (Sireci et al., 1999, p. 306). It appears that these methods are probably most useful for supplementing data derived from other standard setting studies or for evaluating standards already set on a test rather than being standard setting methods in and of themselves.

There are many methods for setting standards as we have briefly reviewed in this section. It appears to be safe to claim that language teachers and testers have little experience in setting cut scores for language tests and are not well aware of how differences in methods affect cut scores. As a result, we are not in a position to select different methods for different purposes. As stated earlier, this area is a critical one, especially when considering the fact that language tests are often used to make high-stakes decisions, which mandates defensible standards. For this reason it is important that studies be conducted to compare standards derived from different methods. Although several previous studies have attempted to set standards, only a few have compared different methods for choosing cut scores. The present study compares three different methods in an attempt to answer the research question, “When these different methods are applied to the same test, do they yield similar cut scores?”

3. METHODS

3.1. Setting

The English as a Second Language Placement Examination (ESLPE) is required of all entering students whose native language is not English and who have not otherwise satisfied their ESL requirement. Results of the ESLPE are used to determine placement into the required sequence of ESL courses or exemption from the ESL requirement. In the case of a non-passing score on the exam, students are placed into one or more of the courses: ESL 33A (Low Intermediate English for Academic Purpose), 33B (Intermediate English for Academic Purposes), 33C (Advanced English for Academic Purposes), and 35 (Approaches to University Writing for ESL students).
3.2. Participants

3.2.1 Raters

All standard setting methods involve judgments, and therefore all require qualified participants. In particular, one of the most important points for defending a set of performance standards is to demonstrate that the panel is substantial in size and representative of the various stakeholders. Judges from two different stakeholder groups participated in the modified Angoff standard setting workshops. The first group of raters was instructors in the ESL program who were familiar with the placement standards and student population. The program had performance standards only for the writing section. In addition, teaching assistants (TAs) are often assigned to teach the same courses across academic years, and thus some of them were not familiar with the performance standards for the courses that they had never taught before. This practice is based on the program’s policy to give TAs a lighter load in preparing for classes they are teaching because these instructors are graduate students as well as TAs. Therefore, only TAs who had taught at least two different course levels and who had served as composition raters were selected in order to ensure that they were familiar with the performance standards. Seven instructors participated in the modified Angoff study, and they were recruited to participate in this study on a voluntary basis.

3.2.2. Test takers

Two groups of test takers participated in the study. For the borderline-group method, four instructors and the ESLPE coordinator identified 23 borderline students and these students took the new ESLPE in computer labs on the campus in 2003. The number of borderline students at each level was 3 for 33B, 8 for 33C, 8 for 35, and 4 for the exempt level. For the cluster class analysis, 116 international students participated in the trial of the new ESLPE. These students were recruited to participate in this study on a voluntary basis. The number of students at each placement level was 4 at the 33B level, 59 at the 33C level, 46 at ESL 35 level, and 9 at the exempt level.

3.3. Materials

This section describes the materials used in the present study.
3.3.1. Test tasks

The present standard setting and validation study was conducted in the context of the new ESLPE at a research university in United States. Since the new ESLPE is being developed as a web-based placement test, the test will be administered in computer labs on the campus. Even though the new ESLPE consists of listening comprehension, reading comprehension, and composition sections, the present study aimed to set standards only on reading and listening sections. The reading section consists of three test tasks: incomplete outline task (Task 1), limited response task (Task 2), and gap-fill task (Task 3). The listening comprehension section consists of two test tasks: the incomplete outline task (Task 4) and the limited response task (Task 5). Finally, the composition section is intended to measure students’ ability to write a formal academic essay. The student responses were scored independently by two instructors in the program. The inter-rater reliability of the ratings was .98.

3.3.2. Rating materials

After the raters were selected, a training session took place to ensure that they used the performance standards in the same way. Each rater was provided with a rating manual, which described the purpose of the study, rating procedures, descriptions of test tasks, descriptions of performance standards, and samples of rating sheets. When all the participating judges became familiar with their tasks, the test booklets, scoring keys, and rating sheets were distributed, and the raters were asked to provide their initial ratings.

3.4. Procedures

This section describes how the cut scores were derived from the three standard setting methods.

3.4.1. Modified Angoff method

In a typical Angoff study, the level of analysis is items, and thus raters estimate the probability that borderline test takers will answer an item correctly. In the present study, however, the level of analysis was test tasks. This decision was based on the recognition of the following problems. First, the test tasks are not discrete-point questions. Since items in each test task are often closely related to each other and discourse-based, raters who participated in the pilot study reported that they found it difficult to make separate ratings for each item. Second, it was not feasible and cognitively too demanding for the raters to
make 284 ratings in the standard setting session (71 items times 4 levels).

The rating sessions consisted of three sessions: training session, Round 1 rating, and Round 2 rating. Since the training of judges is critical to any standard setting application (Reid, 1991), after the raters were selected, a training session took place to ensure that they used the performance standards in the same way. In order to help the raters understand what constituted the minimum level of performance at each placement level, the training of judges involved explaining the purpose of the study and briefly describing the assessment instrument and the five placement levels. Each rater was then provided with rating manuals, which described the purpose of the study, rating procedures, descriptions of test tasks, descriptions of performance standards, and samples of rating sheets. The rating task booklet was then distributed, and each judge was asked to make preliminary judgments for two sample tasks (incomplete outline and limited response) and then to announce their choice of probability for each task. These numbers were then written on a blackboard so that all the judges were able to see them. As suggested by Zieky and Livingston (1977), when the numbers were not all similar, the judge who chose the highest numbers was asked to explain the reasons for choosing high numbers. Then the judge who chose the lowest numbers was asked to explain the reasons for choosing the low numbers.

When all the participating judges became familiar with their tasks, the test booklets, scoring keys, and rating sheets were distributed, and the raters were asked to provide their initial ratings. For the listening section, they were allowed to watch the video-lecture on the web individually. When the Round 1 rating was completed, the raters provided with normative information (Cizek & Bunch, 2007), which shows how their ratings compare with the other participants’ judgments. They were allowed to discuss their initial ratings as a group task by task and then to fine-tune their initial ratings, if necessary. In the Round 2 rating session, reality information (Cizek & Bunch, 2007), that is, the results of pilot tests, was provided so that the raters had a meaningful frame of reference for providing their ratings. Specifically, the test data included item difficulty indices and descriptive statistics (the mean, standard deviation, minimum score, and maximum score) of student performance at each level. After the raters discussed the pilot test results as a group, they were then allowed to make adjustments to their own ratings.

3.4.2. Borderline-group method

In the present study, ESL instructors identified borderline students based both on their experience with them and on their class performance, and these students took the new ESLPE in computer labs on the campus. The median of the distribution of test scores earned by the borderline students at each placement level was set as the cut score at each placement level.
3.4.3. Cluster analysis

Cluster analysis is a statistical method for finding subtypes of related cases from multivariate data. In order to classify test takers into their most likely clusters, test-taker performance data were submitted to a series of cluster models. Hierarchical cluster analyses were first conducted to get a rough idea of the optimal number of clusters to be derived. A series of K-Means analyses were then performed to sort students into groups of pre-specified numbers, and logistic regression was used to identify the optimal cut points between each cluster based on the raw score scale. Even though the sample size for the cluster and logistic regression analyses was not large, it exceeds the minimum sample size required for the two analyses (Formann, 1984; Pampel, 2000).

4. RESULTS

4.1. Modified Angoff Method

In the first Round of rating, seven instructors made judgments for the five test tasks on an individual basis. Overall, the judges experienced more difficulties predicting the expected level of performance at the lower levels than at the upper levels, which led to the greater discrepancy among their ratings for the ESL 33B and 33C. In the group discussion, the raters reported that it was easier to rate “extremes”. The raters also felt more confident with upper level ratings.

After the raters produced their first ratings, they had opportunities to discuss their initial ratings as a group task by task. When they completed the discussion, they were allowed to revise their initial ratings, if necessary. Except for Rater 4, all the raters revised their initial ratings after the group discussion. Of 36 changes in ratings, 30 were lowered whereas six were increased. In the Round 2 rating session, the results of pilot tests were provided, and as a group, they had opportunities to interpret and discuss the pilot test results. The performance data consisted of item statistics and the descriptive statistics of the test tasks at the placement group levels. The judges were allowed to make adjustments after they compared their initial ratings with the pilot test results. The descriptive statistics of the final ratings are reported in Table 1.

The final cut scores on the new ESLPE were computed by calculating the trimmed mean of the instructors’ ratings, a process which simply involved eliminating the highest scores and the lowest scores and then averaging the remaining ratings in the usual way. The trimmed mean was used because of the possibility that using the mean might allow one judge with a very high or very low passing score to have an undue influence on the final
cut scores. The product of the trimmed mean multiplied by the total score was then set as
the cut score. Since section scores are used as a basis for composite scores, the cut scores
for Reading were computed by summing the cut scores for Tasks 1, 2, and 3, whereas the
cut scores for Listening were derived by adding the cut scores for Tasks 4 and 5. The final
cut scores derived from the modified Angoff are presented in Table 2.

4.2. Borderline-Group Method

The borderline-group method requires experts to make judgments about who the
borderline cases are in a student population in order to establish what a typical borderline
performance would be on the test tasks. In this study, five instructors (ESL 33B, ESL 33C
(1, 2), ESL 33C Graduate, and 35) were asked to identify borderline test takers in their
ESL classes whose performance was right around the performance standard under
consideration (i.e., between the two performance standards).

<table>
<thead>
<tr>
<th>Task</th>
<th>Level</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task1</td>
<td>33B</td>
<td>36</td>
<td>60</td>
<td>50.29</td>
<td>10.73</td>
</tr>
<tr>
<td>Task1</td>
<td>33C</td>
<td>55</td>
<td>70</td>
<td>65.00</td>
<td>7.07</td>
</tr>
<tr>
<td>Task1</td>
<td>35</td>
<td>67</td>
<td>80</td>
<td>76.71</td>
<td>4.89</td>
</tr>
<tr>
<td>Task1</td>
<td>Exempt</td>
<td>80</td>
<td>95</td>
<td>89.43</td>
<td>4.54</td>
</tr>
<tr>
<td>Task2</td>
<td>33B</td>
<td>33</td>
<td>65</td>
<td>46.14</td>
<td>13.13</td>
</tr>
<tr>
<td>Task2</td>
<td>33C</td>
<td>50</td>
<td>70</td>
<td>58.43</td>
<td>8.60</td>
</tr>
<tr>
<td>Task2</td>
<td>35</td>
<td>60</td>
<td>80</td>
<td>70.86</td>
<td>6.12</td>
</tr>
<tr>
<td>Task2</td>
<td>Exempt</td>
<td>80</td>
<td>90</td>
<td>82.57</td>
<td>3.82</td>
</tr>
<tr>
<td>Task3</td>
<td>33B</td>
<td>10</td>
<td>30</td>
<td>22.57</td>
<td>6.92</td>
</tr>
<tr>
<td>Task3</td>
<td>33C</td>
<td>30</td>
<td>50</td>
<td>39.57</td>
<td>8.24</td>
</tr>
<tr>
<td>Task3</td>
<td>35</td>
<td>40</td>
<td>65</td>
<td>52.14</td>
<td>9.51</td>
</tr>
<tr>
<td>Task3</td>
<td>Exempt</td>
<td>60</td>
<td>80</td>
<td>68.57</td>
<td>7.48</td>
</tr>
<tr>
<td>Task4</td>
<td>33B</td>
<td>17</td>
<td>30</td>
<td>23.43</td>
<td>6.27</td>
</tr>
<tr>
<td>Task4</td>
<td>33C</td>
<td>30</td>
<td>42</td>
<td>37.14</td>
<td>4.49</td>
</tr>
<tr>
<td>Task4</td>
<td>35</td>
<td>40</td>
<td>65</td>
<td>50.00</td>
<td>8.66</td>
</tr>
<tr>
<td>Task4</td>
<td>Exempt</td>
<td>50</td>
<td>70</td>
<td>64.86</td>
<td>7.44</td>
</tr>
<tr>
<td>Task5</td>
<td>33B</td>
<td>25</td>
<td>55</td>
<td>42.86</td>
<td>9.94</td>
</tr>
<tr>
<td>Task5</td>
<td>33C</td>
<td>44</td>
<td>65</td>
<td>55.57</td>
<td>7.57</td>
</tr>
<tr>
<td>Task5</td>
<td>35</td>
<td>56</td>
<td>80</td>
<td>70.14</td>
<td>7.31</td>
</tr>
<tr>
<td>Task5</td>
<td>Exempt</td>
<td>75</td>
<td>90</td>
<td>80.71</td>
<td>4.50</td>
</tr>
</tbody>
</table>

Based on their judgment of student performance, they identified nineteen students as
borderline cases, and the mean of the performance of these borderline students was set as
the standard of the new ESLPE. For the Exempt level, among the test takers who participated in the trials of the new ESLPE, the ESLPE coordinator first identified students who were exempted from the ESL requirements, and then the bottom four students were selected as the borderline students. The median of the performance of these students was then set as the cut score for the level. The sums of the cut scores for Tasks 1, 2, and 3 were set as the cut scores for Reading, and the sums of the cut scores for Tasks 4 and 5 were established as the cut scores for Listening. The cut scores derived from the borderline-group method are presented in Table 3.

<table>
<thead>
<tr>
<th>Task (Total score)</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Reading section</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Listening section</th>
</tr>
</thead>
<tbody>
<tr>
<td>33B</td>
<td>5.5</td>
<td>4.8</td>
<td>6.6</td>
<td>16.9</td>
<td>2</td>
<td>6.79</td>
<td>8.79</td>
</tr>
<tr>
<td>33C</td>
<td>7.8</td>
<td>6.76</td>
<td>11.7</td>
<td>26.26</td>
<td>3.76</td>
<td>8.96</td>
<td>12.72</td>
</tr>
<tr>
<td>35</td>
<td>8.25</td>
<td>8.54</td>
<td>16.5</td>
<td>33.29</td>
<td>5.13</td>
<td>11.36</td>
<td>16.49</td>
</tr>
<tr>
<td>Exempt</td>
<td>9.92</td>
<td>10.08</td>
<td>21</td>
<td>41</td>
<td>6.47</td>
<td>12.16</td>
<td>18.63</td>
</tr>
</tbody>
</table>

4.3. Cluster Analysis

This standard setting approach is different from the other standard setting methods because it analyzes test score data rather than raters’ judgments. Since cluster analysis involves statistical analysis of actual test score data, it is expected to supplement data derived from the modified Angoff method and borderline-group method. The data for all the participants were analyzed (N = 116). Seven students who had missing data were eliminated from the analysis because these scores could have adversely affected the cluster solutions. Therefore, the final sample comprised 109 students. The subscores derived from items comprising each section were used as the two variables to obtain clusters. After preparing the data by computing the subscores for Reading and Listening, the data sample was first analyzed using hierarchical cluster analyses to get a rough idea of the optimal number of clusters to be derived. The optimal cluster solutions
were identified by interpreting the change in the agglomeration coefficient. The results of this process show that large jumps occurred between the third and second cluster solutions, indicating that the three-cluster solution appeared viable in the sample. Based on the results, it seemed sensible to explore 2-4 cluster solutions for the K-Means analyses.

A series of K-Means analyses were performed to sort students into groups of pre-specified numbers, and an index of internal cohesion was used in selecting the optimal clustering solution. Another criterion for evaluating the solution was stability across the samples in terms of cluster profiles. A three-cluster solution was judged to be the optimal partitioning of the data for the purpose of setting standards on the test. Descriptive statistics of the final cluster solution are presented in Table 4. The first column lists the cluster and the second column gives the proportion of the 109 students in each cluster. The next two columns give the mean test scores for the students in each cluster.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>N</th>
<th>Reading</th>
<th>Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>24 (22%)</td>
<td>39.71</td>
<td>18.75</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>37 (34%)</td>
<td>20.43</td>
<td>10.46</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>48 (44%)</td>
<td>31.00</td>
<td>13.27</td>
</tr>
</tbody>
</table>

The first cluster constituted 22% of the test takers, and the second and third clusters comprised 34% and 44% of the students, respectively. The mean performance of the three clusters differs more in the Reading section than in the Listening section.

Since cluster analysis forms clusters even when no true clusters exist in the data, one-way Analyses of Variance were run on the K-Means solution to determine if the clusters were statistically different from each other in the same solution. The results show that all clusters were significantly different for both reading and listening at the .05 level of significance. The results of the ANOVA are provided in Table 5.

The Tukey’s b post-hoc comparison was also carried out to examine whether all pairwise comparisons between clusters were significant. The Tukey’s b procedure was chosen because it is appropriate when sample sizes are not equal for all comparisons. The results show that all pairwise comparisons were statistically significant in terms of the cluster members’ performance on the reading and listening sections.
The cut scores between the clusters were derived from the cluster analysis solutions in two different ways. First, as Livingston and Zieky (1989) suggest, since test scores from test takers who belonged to two clusters overlapped, logistic regression was used to identify the cut score that was associated with a .5 probability of a student being classified as an upper- or lower- group member. In the logistic equation, the dependent variable was dichotomous - membership in one of two clusters - and the covariate was the total raw score for each examinee. The following equation was used to define the conditional probability function, and logistic regression was used to estimate the a and b parameters in the equation.

\[ P = \frac{1}{1 + e^{-(a+bx)}} \]

Cut scores were then derived by setting \( p = .5 \), and solving for \( x \). The results of the logistic regression analyses are presented in Tables 6, 7 and 8.

### TABLE 6

<table>
<thead>
<tr>
<th>Score</th>
<th>Constant</th>
<th>1.481</th>
<th>5.23</th>
<th>4.399</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig.</td>
<td></td>
<td>3.99</td>
<td></td>
<td>.005</td>
</tr>
<tr>
<td>Exp(β)</td>
<td></td>
<td></td>
<td></td>
<td>4.399</td>
</tr>
</tbody>
</table>

### TABLE 7

<table>
<thead>
<tr>
<th>Score</th>
<th>-8.552</th>
<th>1.982</th>
<th>1.00</th>
<th>0.000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
</tr>
</tbody>
</table>

### TABLE 8

<table>
<thead>
<tr>
<th>Score</th>
<th>-.54188</th>
<th>19.212</th>
<th>1</th>
<th>.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td></td>
<td></td>
<td></td>
<td>0.000</td>
</tr>
</tbody>
</table>
Second, since test scores in Reading section from test takers who belonged to the clusters 1 and 2 did not overlap, the cut score was set on the score halfway between the minimum score of the cluster 1 and the maximum score of the cluster 2, which was 26. The derived cut scores are presented in Table 9.

<table>
<thead>
<tr>
<th>Course</th>
<th>Reading</th>
<th>Listening</th>
</tr>
</thead>
<tbody>
<tr>
<td>33B</td>
<td>26</td>
<td>11</td>
</tr>
<tr>
<td>33C</td>
<td>36</td>
<td>17</td>
</tr>
<tr>
<td>Exempt</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The final three-cluster solution did not correspond to the five placement levels, which may be explained by the limitations of the data set. First, the number of students at the level of 33B and Exempt were relatively small (4 and 9, respectively). In addition, some of the students in 33C and 35 students were originally placed into 33B and 33C and then promoted to these levels. As Brown (1980) shows, newly placed students typically perform better than these continuing students within the same classroom. It also needs to be pointed out that the results are based on listening and reading scores only because the students did not take the writing exam. It may be possible that writing scores would have helped to distinguish the groups.

4.4. Comparison of Cut Scores From the Three Methods

One of the driving questions of the present study was whether the three standard setting methods set cut scores on the test at the same place. Whereas both the modified Angoff method and borderline-group method produced the four cut scores, the cluster analysis yielded only two cut scores. Table 10 shows the cut scores derived from the three standard setting methods, rounded to the nearest whole number.

Overall, the borderline-group method tended to produce relatively lower cut scores than the modified Angoff method and the cluster analysis. The modified Angoff method and the borderline-group method yielded the same cut score for the ESL 33B level in both Reading and Listening and a very similar cut score for the ESL 33C level in both sections. However,
the cut scores appear to differ at the ESL 35 and Exempt levels, with the modified Angoff method having higher cut scores than the borderline-group method, in particular, in Reading. However, the two methods were in close agreement with respect to where the cut score point for the Exempt level in the Listening section should be established. The cluster analysis produced cut scores similar to those produced by the other two methods for ESL 33C in both Reading and Listening and ESL 35 in Listening. However, it yielded a much higher estimate for ESL 35 in Reading. The cluster analysis and the borderline-group method differed most in where they set the cut scores for the ESL 35 levels in Reading. The score difference was 7.

5. DISCUSSION AND CONCLUSIONS

This study investigated how the ESL program set cut-scores on its new placement exam. As described earlier, the cut scores derived from the three standard setting methods did not entirely agree with each other in assigning students into different levels. The modified Angoff method and the borderline-group method produced very similar cut scores for the lower levels, that is, ESL 33B and 33C, whereas they differed in where they established cut scores for the higher levels, that is, ESL 35 and Exempt, with the modified Angoff method producing higher cut scores than the borderline-group method. The cluster analysis yielded two cut scores, which were very similar to the cut scores derived from the modified Angoff method at the corresponding levels. The close convergence is quite encouraging in that the judges’ ratings were supported by the statistical groupings based on actual student responses. The borderline-group method tended to produce lower cut scores than the other two methods. The results also indicate that the cluster analysis provides additional information that can be useful for helping set standards on language tests.

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Listening</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angoff</td>
<td>Borderline</td>
<td>Cluster</td>
</tr>
<tr>
<td>33B</td>
<td>17</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>33C</td>
<td>26</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>35</td>
<td>33</td>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td>Exempt</td>
<td>41</td>
<td>37</td>
<td>19</td>
</tr>
</tbody>
</table>

The results of the study have shown that the choice of standard setting method influenced the resulting cut scores. The reason for the discrepancy may be accounted for by the following issues. First, partial credit scoring was allowed for Tasks 2 and 5. In
particular, Task 5 had a 3-point item and another 2-point item. Since test task was the unit of analysis in the modified Angoff method, the raters were not able to take the scoring scheme into consideration when they produced ratings. This may explain why the modified Angoff method tended to yield higher cut scores for Tasks 2 and 5 across the placement levels. It also needs to be pointed out that some raters experienced predicting the performance of minimally competent students because of the lack of performance standards. As a result, it is possible that they overestimated the minimally acceptable level of performance for each level.

Second, only three 33B students participated in the trials of the ESLPE. Perhaps due to the small sample size, ESL 33B did not emerge as a separate group in the cluster analysis, with three 33B students being grouped with cluster 2. The same is true for the Exempt group in that the data included only nine students at the Exempt level. It is, therefore, possible that some of the borderline students may not belong in the borderline-group of the level at which they were placed and thus the standards may not represent a typical borderline performance.

At the same time, it is also possible that false negative classification errors could have been made when the instructors identified some of the borderline students based on their Fall quarter test scores. Since the present study was conducted at the end of Winter quarter, the students may have already been exposed to academic English in their regular courses. As a result, it is possible that some of the borderline students no longer belonged in the borderline-groups. In addition, the size of the group judged to be on the borderline was relatively small, which could have made the estimates of cut scores unstable (Cizek & Bunch, 2007).

As the findings of this study show, however, it is a challenging task to establish defensible cut scores on language tests. Perhaps the most controversial problem in standard setting today is the fact that there is no agreement as to a best method and that different standard setting methods would give different results. Unfortunately, the measurement literature does not agree on what to do about the problem. How does one determine which standard is right? If three methods yield three different results, are all of them wrong? Is one of them right? Is it perhaps the one that passes more test takers? Or is it the one that minimizes false positive errors in classification? The only widely recognized answer to these questions is to choose standard setting methods, which are most appropriate for a particular situation and purpose. Therefore, the process always entails human judgment, and this fact “is probably the single greatest area of agreement in the history of setting standards” (Zieky, 1995, p. 28).

Even though standards are often the basis for making important decisions that may strongly affect test takers’ lives, standards are not commonly discussed in the language testing field. Standard setting is often left to the last stage of the test development process.
and, as a result, carried out haphazardly, which can result in standards that are not as useful as they might be. In situations where classification decisions are to be made on the basis of test scores, defining performance standards and setting cut scores should constitute an essential component of the test development process, and test standards should be a concern at all test development stages. Therefore, it is important that test developers and users have a formal plan so that standard setting is not just an afterthought nor carried out arbitrarily. The plan should specify the various points at which they will collect evidence regarding the usefulness of the standards.

Given that standards are often the basis for making important decisions that may dramatically affect our students, future studies should also examine the impact of test standards on test takers, programs, and society. For example, language testers and users could try to follow up on people who failed and/or passed the test to determine how many of these test takers were false negatives or false positives, along with the reasons for why they may have either over-performed or underperformed on the test. Research questions may include: Is there evidence that many of them were actually qualified to pass at the time they took the test? Is there evidence that many of the people who passed the test were unqualified? What were the consequences of failing a qualified person or passing an unqualified one?

Future studies could compare final placement results with classifications made on standard setting methods. The results of this comparison would reveal how many false-positive and false-negative misplacements would result if placement decisions were made based on the cut scores derived from each method. Future studies could also examine how the impact data, which tells how many students would have been placed into each level, influences the determination of the cut scores.

Since the use of standard-based tests are on the rise for accountability purposes, future studies should also explore the impact of these high-stakes standard-based tests on students, classroom instruction and society. The introduction of a standard-based approach has, on occasions, proven problematic for a variety of political, technical, and practical reasons (Brindley, 2001). Given the high stakes that are frequently associated with standard-based assessment, it is important to investigate the ways in which these problems are being manifested and addressed in the real world. Finally, future standard-setting studies may examine participants’ thought processes and experiences during judgmental standard-setting processes.

Since this standard setting study was conducted in the context of the new institutional placement at a research university, the results of this study may not be generalizable to other contexts. As Livingston and Zieky (1982) point out, there is no one method that is best for all testing situations, and thus findings from this study need to be interpreted in the context of this particular program. Even though it is true that standard setting procedures...
do need to be tailored to the specific concerns of individual language programs, it is hoped that the standard setting procedures described in this present study are a starting point for future standard setting studies that constantly reshape and redefine themselves, depending on the context of the test and the program. The judicious selection of standard setting methods, sound implementation of the methods, and evaluation of the usefulness of derived standards will increase considerably the likelihood of producing defensible and valid standards so that language tests can achieve their intended goals. Language testers and test users must use the best available techniques currently available to establish defensible standards so that high stakes decisions are not being made unsystematically, covertly, and perhaps, unfairly.

It is true that standard setting is difficult and perhaps arbitrary in nature (Glass, 1978). It is clearly not as exact as we would like. However, one thing seems clear from the discussion heretofore. Standards are here to stay, and well-considered standards are better than no standards at all. Therefore, language testers and test users must use the best available techniques currently available to establish defensible standards so that high stakes decisions are not being made unsystematically, covertly, and perhaps, unfairly. It is hoped that these findings will contribute to ongoing research investigating the validity of performance standards and cut-scores on language tests.

REFERENCES

How Good is Good Enough?: A Comparison of Three Methods for Establishing Cut Scores on ...

Washington, D C: TESOL.


Applicable levels: Elementary, secondary, tertiary

Sang-Keun Shin
Department of English Education
Ewha Womans University
52 Ewhaebi-gil, Seodaemun-gu
Seoul 120-750, Korea
Phone: 02-3277-4648
Fax: 02-3277-2647
Email: sangshin@ewha.ac.kr

Received in September 2014
Reviewed in October 2014
Revised version received in November 2014