Using the results of multiple standard setting methods to inform policy

Susan L. Davis-Becker

Chad W. Buckendahl

Alpine Testing Solutions

Paper presented at the 2011 annual meeting of the National Council on Measurement in Education, New Orleans, LA.

All inquiries to this paper should be directed to Susan.DavisBecker@Alpinetesting.com

Abstract

The availability of external validity evidence to evaluate the results of standard setting processes (Kane, 2001) may be limited depending on the nature of the testing program. In this paper, we evaluate strategies for using the results of multiple standard setting methods as one source of external evidence (Green, Trimble, & Lewis, 2003; Jaeger, 1989) and in settings where other national or international benchmarks may not be available to guide policymakers. Although we expect that different methods would yield different results, practitioners are often asked to provide interpretative guidance to policymakers who make the final cut score decisions. However, to date, standard setting literature has provided little guidance for practitioners in this regard.

Our approach to providing this guidance is to evaluate the process and results of each standard setting method applied and then use the relative measure of strength to weight the results of the standard setting process. We illustrate this concept by comparing and contrasting case studies from two different testing programs. The first example was from a school district graduation eligibility testing program for which there was not an addition relevant, external assessment available to inform the policy decision. The second example was from an international licensure program that utilized multiple test-based standard setting methods to recommend a cut score. We conclude by discussing strategies for communicating such results to policymakers.

Using the results of multiple standard setting methods to inform policy

Standard setting is often one of the most challenged and controversial steps in the test development process. Conducted properly, it requires the use of a defensible method, the judgments of an appropriate sample of subject matter experts, the consideration of key factors (e.g., intended use of test scores, impact of cut score, consequences of test performance), and is ultimately a decision made by policy makers.

The impact of a cut score can be significant depending on the nature of a program. In education assessments, cut scores are used for multiple purposes such as classifying exam performance of students for federal accountability requirements and determining if students meet requirements for graduation. The results can have an impact at the student, teacher, school, district, and state levels. In certification and licensure (collectively, credentialing), cut scores are used to classify exam performance as acceptable for entry level practice in a profession or less than acceptable. The results of these credentialing assessments can have an impact at the candidate, training program, and national or international level depending on the nature of the program.

Given the scope of impact that a cut score can have, it is important that those responsible for making the final cut score decisions make an informed decision using all relevant information. The foremost factor should be the results of some type of formal standard setting procedure. There are numerous methods discussed in the literature that have been defined, modified, reviewed, evaluated, and ultimately supported by the psychometric community (see Hambleton & Pitoniak, 2006 for a review of current methods). Rather than prioritizing one method over another, researchers have taken the approach of evaluating the fit of a standard setting method to a situation (test, use of scores, resources), the way in which the process is conducted, and the evidence available to support the application of the method (e.g., Hambleton, 2001; Kane 1994; 2001).

Within the evaluative frameworks that have been published on standard setting, one recommendation is to consider the results of multiple standard setting methods. However, research has consistency shown that that different standard setting methods will likely yield different results (e.g., Jaeger, 1989; Kane, 2001) and that there is not a true cut score (e.g., Jaeger, 1989; Zieky, 2001). So how does one resolve discrepant results across standard setting methods? In this paper we attempt to provide some initial guidance for resolving differences that emerge when using multiple standard setting methodologies. Further, we propose strategies for practitioners in helping policymakers interpret the potentially divergent results of standard setting information.

Evaluation of Standard Setting Procedures

Kane (2001) suggested three sources of validity evidence to consider when evaluating the results of a standard setting process: procedural, internal, and external. Most evaluative criteria proposed by standard setting researchers ties into one of these three areas. For procedural validity evidence, practitioners should evaluate the selection and execution of the standard setting method. This includes participant selection and qualifications, appropriateness of methodology, the application of the methodology, and the participants’ perspectives about the implementation of the methodology as a few primary sources of procedural evidence. Internal validity evidence is provided by the consistency of participants' ratings (e.g., correlation) and the convergence of participant's recommended cut scores (e.g., measure of variability). Finally, external validity evidence comes from triangulating the results of the standard setting process with some other indicator of where the cut score should be that is external to the process or another measure of ability by the same examinee population. Any weaknesses in the validity argument that are observed in these areas can be used to mitigate policymakers' judgments regarding the usefulness of the participants' recommendations and the validity of the interpretation.

The type of evidence that Kane refers to as external is generally the most difficult to collect as it requires the use of an additional panel of experts, an additional method, or obtaining results from another measure of the same construct. In the latter option, practitioners must find data for a second exam taken by the target population that is designed to measure the same construct at the same level with cut scores that were set with the same expectations. Kane (2001) cautions that this second assessment "should not be chosen simply for convenience" (p.76). Ideally, programs could target a higher level (e.g., district to state, national to international) or parallel level examination as this second source of validity evidence. For educational programs at all levels (district, state, regional, national, and international), there is often a lack of a parallel examinations, particularly in the case of developing nations that may not participate in some of the international assessments (e.g., TIMSS, PISA, PIRLS). The same challenge is faced by credentialing programs across borders. Therefore, programs turn to the second option of multiple standard setting methods.

For those programs with the resources available to conduct multiple standard setting procedures there still remains the challenge of properly utilizing the results. Cizek and Bunch (2007) express concern about the lack of research or guidance that exists to guide interpretation of the results of multiple standard setting methods. Too often, policymakers are faced with the task of weighing each piece of information in making the final cut score decision. Geisinger (1991) and Geisinger and McCormick (2010) identified ten considerations for those making decisions on cut scores, but stop short of providing specific guidance for how practitioners should use these to assist policymakers. With regards to the results from multiple standard settings, Geisinger and McCormick (2010) recommend "Decision makers should be prepared to combine results from these different sessions if their results should diverge, perhaps with a predetermined weighting system" (p. 42). Although we agree that policymakers should anticipate a range of results when using multiple methods, we feel it is the responsibility of the practitioners designing and conducting the standard setting processes to provide interpretive guidance. The suggestion to use a predetermined weighting system as part of this process assumes a priori knowledge of the validity of each source prior to data collection.

Cizek and Bunch (2007) express an opinion that there is likely a best standard setting method for a given application and therefore practitioners should focus on identifying and using that method. We contend that there is no best method but that the use of a particular method can be better than the use of another method for a given situation i.e., the match of the method to the situation would be one factor in determining the success of the application. Further, that the combination of methodologies from different perspectives (e.g., test-centered, examinee-centered) may yield better information for policymakers in their decision-making process.

Guidelines for Interpreting the Results of Multiple Standard Setting Methods

Within the research on standard setting, there are several sets of guidelines and frameworks for evaluating standard setting processes (e.g., AERA, APA, & NCME, 1999; Cizek, 1996; Kane 1994; 2001; Hambleton, 2001). In a review of this research, one finds a general level of agreement among the authors as evidenced by the consistency of the evaluation criteria. In our review of these criteria, we attempted to identify the specific elements that one could use to evaluate the strength and weakness of different methodologies as applied to a given examination.

The result of this review is a list of 15 criteria that practitioners could apply in a situation where divergent results exist across multiple standard setting methods (see Table 1). An important caveat that accompanies this table is the understanding that not all criteria should be interpreted as being of equal importance. For example, suppose two methods were used to set a cut score. The only weakness identified for method # 1 is that no evaluation data was collected to support the panelists' understanding of the methodology and confidence in their judgment (criterion #10) whereas in method #2 the only weakness was that the panelists were not qualified (criterion #1). These two weaknesses are not equal. Criterion #1 (qualified panelists) is of fundamental importance. Without it there would be no confidence in any results.

Table 1. Evaluative criteria for comparing the merits of multiple standard setting methods

Criteria	**Sources for Recommendation**[1]
Procedural	
1. Were the panelists qualified to make the judgments that were asked of them?	*Standard* 1.7, Hambleton, 2001; Kane, 2001
2. Was the panel representative of the population who should be involved?	*Standard* 1.7, Hambleton, 2001; Kane, 2001
3. Did the panelists have a good understanding of the test, the purpose of the test, the intended use of scores, and the abilities of the intended population?	Cizek, 1996; Hambleton, 2001; Kane, 2001
4. Did the panelists have a good understanding of the performance levels for which they were setting cut scores?	*Standard* 4.21, Cizek, 1996; Hambleton, 2001; Kane, 2001
5. Were the panelists properly trained on the standard setting method and prepared to make the judgments that were asked of them?	*Standard* 4.21, Cizek, 1996; Hambleton, 2001
6. Was the standard setting method appropriate for the test, the intended use of scores, and the panelists?	Cizek, 1996; Cizek & Bunch, 2007; Hambleton, 2001
7. Was the standard setting method carried out in the way in which it was designed – or were any adjustments justified and appropriate?	Cizek & Bunch, 2007
8. Was an iterative processes used so that panelists were able to review their recommended cut scores before they were finalized?	Kane, 2001
9. Were the panelists provided data as input to their decisions (e.g., impact, item difficulty)?	*Standard* 4.21, Hambleton, 2001; Kane, 2001
10. Does the evaluation data support the panelists' understanding of the methodology and their confidence in their judgments?	Cizek, 1996; Cizek & Bunch, 2007; Hambleton, 2001; Kane, 2001
Internal	
11. Do the ratings (recommended cut score and item-level if applicable) suggest the panelists were in general agreement with one another?	Kane, 2001
12. Do the panelists' item-level ratings correlate with available measures of item difficulty? *(for test-based methods)*	Kane, 2001
13. Do the panelists' examinee-level ratings correlate with other measures of examinee ability? *(for examinee-based methods)*	*Standard 4.20,* Kane, 2001
External	
14. If performance level expectations were developed during the standard setting – are they reasonable and consistent with other documented expectations?	Cizek & Bunch, 2007
15. If applied, would the recommended cut scores result in a reasonable classification of examinees into performance categories?	Cizek & Bunch, 2007; Kane, 2001

To exemplify how these criteria could be applied, we next describe two case studies that were adapted from actual standard setting processes to protect the confidentiality of the programs.

[1] Sources provided are examples of where readers can find additional information about each criterion. This is not mean to be an exhaustive list.

Applying the Evaluation Framework

Educational District-level Assessment

Our first example focuses on the standard setting process for a grade-level mathematics assessment within a school district. The district was responsible for developing, administering, and reporting on the local mathematics achievement of their students. At the time the performance standards were set, the district did not have a state assessment as a means of external evidence for comparison and realized the intended use of their district mathematics assessment was divergent from any other assessment their students participated in (e.g., NAEP). Therefore, they obtained external validity evidence by using multiple standard setting methods: a variation of the Hofstee (1983) method, a modified Angoff (1971) method, and a contrasting groups method (Livingston & Zieky, 1982). Participants in the study were educators from across the district. The standard setting study began with an orientation that covered the purpose of the meeting and an overview of the standard setting methods to be used. This was followed by a discussion of the four performance levels as defined by the district and a discussion of the transition points between the performance levels which defined where the three cut scores were should be identified.

Panelists began providing judgments with the variation on the Hofstee (1983) method. Specifically, they were asked to estimate what percentage of students in the district, would be in each of the four performance categories (Levels 1, 2, 3, and 4). These overall percentages were then compared to the performance of all students in the district and cut scores were identified that, if applied, would result in percentages of students across the four performance categories that would closely approximate the average percentages estimated by the panelists. This modification represents only part of the full set of judgments that are included in the actual Hofstee (1983) method but was used to gain the panelists' initial estimates on the distribution of students across the four performance categories.

For the second method in this study, panelists participated in a modified Angoff method where they were asked to make Yes/No judgments (Impara & Plake, 1997) on each item for students at the transition point in each performance level (i.e., do you think a student who is barely level 2 would answer this item correctly? Barely level 3? Barely level 4?). After completing their initial judgments, panelists received feedback that include their initial recommended cut scores, the panel's initial recommended cut scores, the impact if the panel's initial cut scores were used (i.e., % of students in each category) and empirical estimates of item difficulty (i.e., p-values). After a brief discussion of this feedback, panelists were allowed to revise their ratings. The second round of ratings was used to compute the final recommended cut scores.

For the third method, the Contrasting Groups method described by Livingston and Zieky (1982), each panelist was provided a copy of their class roster and asked to indicate the performance level exhibited by each student throughout the school year. To compute the recommended cut scores, the average test score of all students classified within each performance category was estimated, and the median point between the average scores of adjacent groups was estimated as the cut score.

Across these three methods, the modified Hofstee method and Contrasting Groups would be considered examinee-centered and the Angoff would be considered test-centered. Given that these represent such different perspectives, one would expect a difference in the results. Figure 1 shows the recommended cut score, for each of the three higher performance levels (i.e., no level 1 cut score) by method. As is shown in Figure 1, the most notable differences were in the recommended cut scores for Levels 2 and 4.

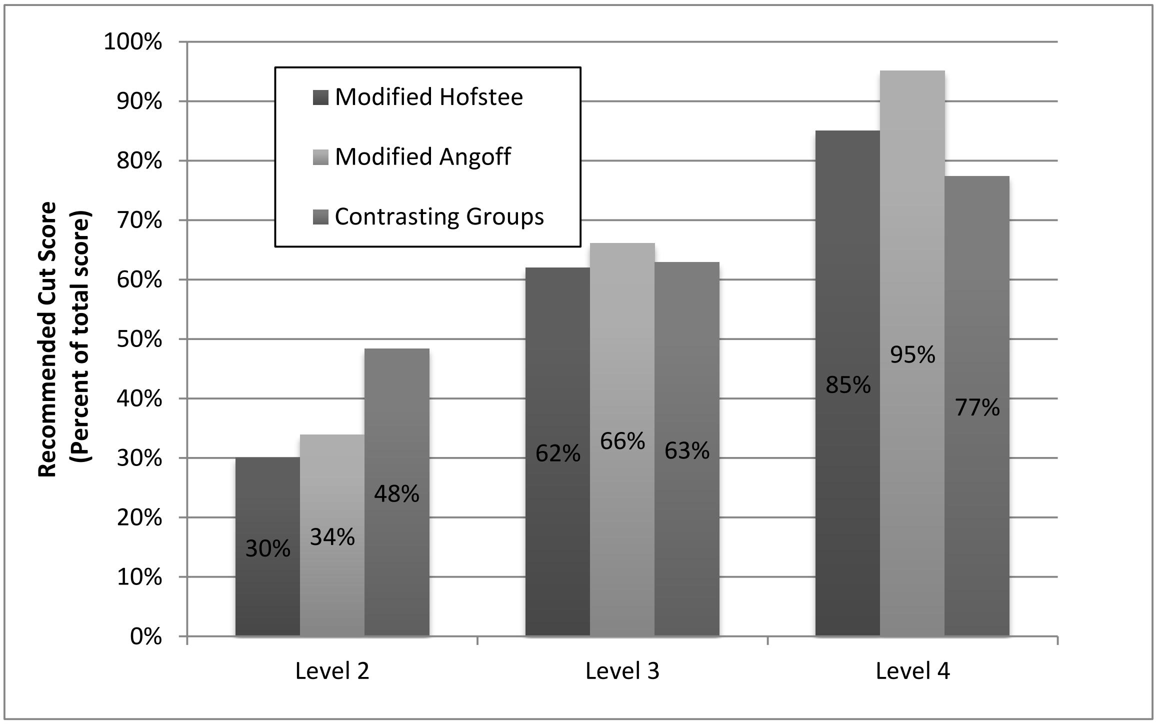

Figure 1. Educational assessment example – comparison of recommended cut scores by method

Another way to evaluate these differences is by considering the impact each set of recommended cut scores would have in terms of performance level classification. This comparative analysis is shown in Figure 2.

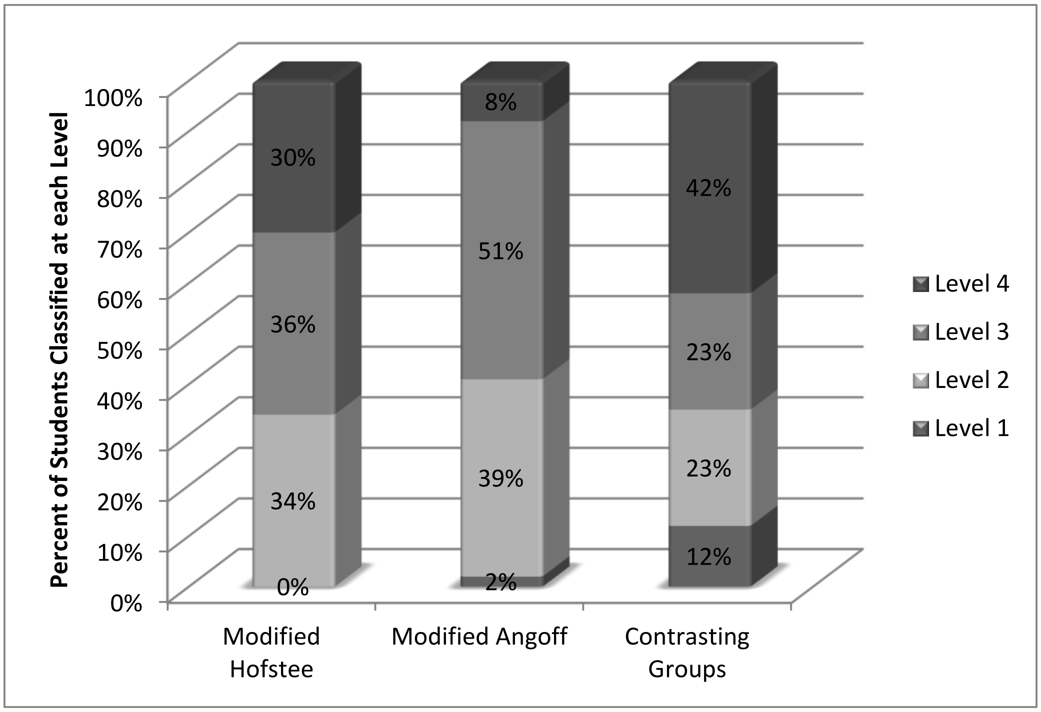

Figure 2. Educational assessment example – comparison of impact by method

Using our evaluative framework, we can identify the validity evidence available to support the outcomes of either of these methodologies. Table 2 summarizes the review of each application based on the aforementioned evaluative criteria.

The purpose of analyzing the methodologies in this way is not to rule them in or out but rather to understand the relative strength of each method and the utility of the results as upper and lower bounds of recommendations to potentially inform the policymaking process. Through the analysis presented in Table 2, it is apparent that the results of the modified Hofstee method have minimal supportive validity evidence compared to the modified Angoff or contrasting groups methods because they do not likely represent the abilities of the student population (e.g., no students at Level 1).

Also from the table, there is concern over the panelists' estimation of how well the level 2 candidates will perform on the exam (Angoff method) as well as lower correlations observed for the levels 2 and 4 item-level ratings due to ceiling and floor effects. On the other hand, a regression towards the mean was observed for the contrasting groups' results for the Levels 2 and 4 cut scores. Therefore, for the Level 2 cut score, if the results of the remaining methods are judged as equally credible, practitioners should advise policymakers to consider the range of recommended cut scores resulting from the two methods with substantial validity evidence (34% - 48%) and to focus on the middle of this range which would balance for the lower bound estimation of student ability through the Angoff method and misclassification of Level 2 students as Level 1. For the Level 3 cut score, the recommended range was much smaller (62% - 66%) and the preponderance of evidence suggests the higher cut scores (Contrasting groups: 63%, Angoff: 66%) are supported by validity evidence. It is important to remember that in such a case, the similar results support the reliability of the results rather than the validity. Finally, for the

Level 4 cut score, the low end of the range (77%) was possibly a result of by over-classification of students’ abilities at Level 4 where some of them should be at Level 3. The high end of the range (95%) is possibly due, at least in part, to a ceiling effect produced by the Angoff methodology when applied to multiple cut scores. Therefore, practitioners should be encouraged to target the middle of the range for identifying the appropriate cut score.

Table 2. Educational Assessment Example – Comparative Analysis of Validity Evidence

Criteria	Modified Hofstee	Modified Angoff	Contrasting Groups	Strengths or weaknesses of each method.
Procedural				
1. Were the panelists qualified to make the judgments that were asked of them?	Yes	Yes	Yes	
2. Was the panel representative of the population who should be involved?	Yes	Yes	Yes	
3. Did the panelists have a good understanding of the test, the purpose of the test, the intended use of scores, and the abilities of the intended population?	Yes	Yes	Yes	Given how the results of the test would reflect on the panelists, they may have been motivated to lower the cut score for the “proficient” performance level (level 2) which they could do through the Angoff method.
4. Did the panelists have a good understanding of the performance levels for which they were recommending cut scores?	Yes	Yes	Yes	
5. Were the panelists properly trained on the standard setting method and prepared to make the judgments that were asked of them?	No?	Yes	Yes	More training was provided for the Angoff method as compared to other methods.
6. Was the standard setting method appropriate for the test, the intended use of scores, and the panelists?	No	Yes	Yes	Angoff and Contrasting Groups allow panelists to make smaller specific judgments whereas Hofstee requires a more holistic judgment.
7. Was the standard setting method carried out in the way in which it was designed – or were any adjustments justified and appropriate?	No	Yes	Yes	The Hofstee was modified to exclude the test score component of the judgments.
8. Was an iterative processes used so that panelists were able to review their recommended cut scores before they were finalized?	No	Yes	No	The Hofstee and Contrasting Groups methodologies, as implemented, were not iterative.
9. Were the panelists provided data as input to their decisions (e.g., impact, item difficulty)?	No	Yes	No	The Hofstee and Contrasting Groups methodologies, as implemented, did not utilize data as an input to the process.
10. Does the evaluation data support the panelists’ understanding of the methodology and their confidence in their judgments?	Yes	Yes	Yes	
Internal				
11. Do the ratings (recommended cut score and item-level if applicable) suggest the panelists were in general agreement with one another?	Yes: Lvls 2 & 3, No: Lvl 4	Yes	Yes	High standard deviation observed for the Level 4 cut score in the Hofstee method.
12. Do the panelists’ item-level ratings correlate with available	N/A	Yes	N/A	Hofstee and contrasting groups methodologies do

Criteria	**Modified Hofstee**	**Modified Angoff**	**Contrasting Groups**	**Strengths or weaknesses of each method.**
measures of item difficulty? *(for test-based methods)*				not produce item-level ratings. Angoff ratings for Level 3 correlated much higher than Levels 2 or 4 due to celling and floor effects.
13. Do the panelists' examinee-level ratings correlate with other measures of examinee ability? *(for examinee-based methods)*	N/A	N/A	Yes	Hofstee and Angoff methodologies do not produce examinee-level ratings. The contrasting groups results panelists identified students as Levels 1 and 4 that are more similar to the students in Levels 2 and 3 (regression towards the mean).
External				
14. If performance level expectations were developed during the standard setting – are they reasonable and consistent with other documented expectations?	Yes	Yes	Yes	
15. If applied, would the recommended cut scores result in a reasonable classification of examinees into performance categories?	No	Yes	Yes	The results of the Hofstee methodologies suggested that no students would be Level 1 (i.e., Below Proficient).

National Licensure Program

Our second example is from a national licensure program. Given that this is a national examination, there is no means for comparison in terms of similar examination that candidates are subjected to. The exam population comes from different training programs and other measures targeting the same constructs would vary by institution. Therefore policymakers chose to implement two standard setting methodologies – modified Angoff (1971; Impara & Plake, 1997) and Bookmark (Mitzel, Lewis, Patz & Green, 2001) – to obtain external validity evidence to support their choice of cut score. The panelists for this study included 40 educators, recently licensed practitioners, and more experienced practitioners.

The workshop began with an overall training whereby the panelists were introduced to the standard setting process, the purpose of the meeting, and both of the methods they would be using in the workshop. As a part of the training, panelists were introduced to the idea of the "minimally competent candidate" (MCC, i.e. target examinee) or a candidate who had the minimum knowledge and skills necessary for licensure in this field. The panel worked together to detail the expectations for the MCC in terms of knowledge, skills, and abilities. In the next component of training, panelists were given an opportunity to practice both the Angoff and Bookmark methods. After the training, panelists were split into two panels – one panel completed the Angoff method first followed by the Bookmark and the other completed the Bookmark method first followed by the modified Angoff method.

For the modified Angoff method, following the training activity, panelists were asked to indicate, for each item, whether the minimally competent candidate they had in mind would answer the item correctly or not (Right or Wrong). After making their initial ratings, panelists were given feedback on the initial results (their recommended passing score, the panel's recommended passing score, p-values by item, answer key for each item, and impact of panel passing score) on the results of the first round. After seeing the data, panelists were allowed to make a second round of estimates which were used to calculate the final recommended cut score.

In our application of the Bookmark method, panelists were presented with an Ordered Item Booklet (OIB) and asked to begin with the easiest item (first page of the OIB) and move through the booklet and identify the place in the booklet that separated the items the MCC would likely (67% probability) of answering correctly (before the bookmark) and the items the MCC would not likely answer correctly (after the bookmark). After the initial bookmark placement, the panelists were given feedback data on the initial results (their recommended passing score, the panel's recommended passing score, p-values by item, answer key for each item, and impact of panel passing score). After seeing the feedback data, panelists were allowed to modify their bookmark placement for how they expected the MCC to perform. The second bookmark could be either the same or different from their initial bookmark. As with the Angoff method, the feedback data provided a reality check to ensure that the recommendation is not unrealistically high or low because the panelist has misjudged the difficulty of the items.

Both methods would be considered test-based as the panelists focused their judgments on the test and item-level. Figure 3 shows the recommended cut scores that estimated based on the second round of ratings for each method. As shown in the Figure, the two recommended cut scores differed by four percent of the total test score.

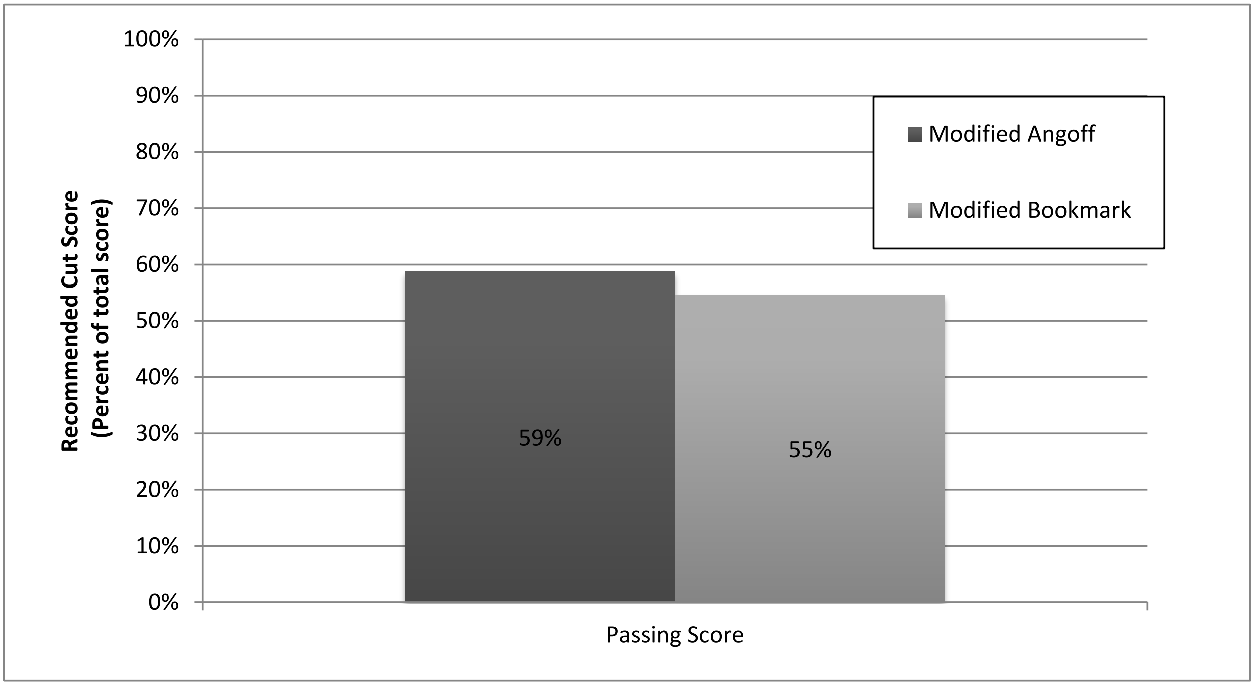

Figure 3. Licensure example – comparison of recommended cut score by method

An additional way to compare the results of the two methods and evaluate the significance of the difference in the recommended cuts cores is to compare the impact on the pass rate. Figure 4 shows the anticipated pass rate if each cut score was implemented in this licensure program. Given that this is a national exam, this 12% difference is substantial in terms of the number of candidates that would be affected by where the cut scores was set within this range.

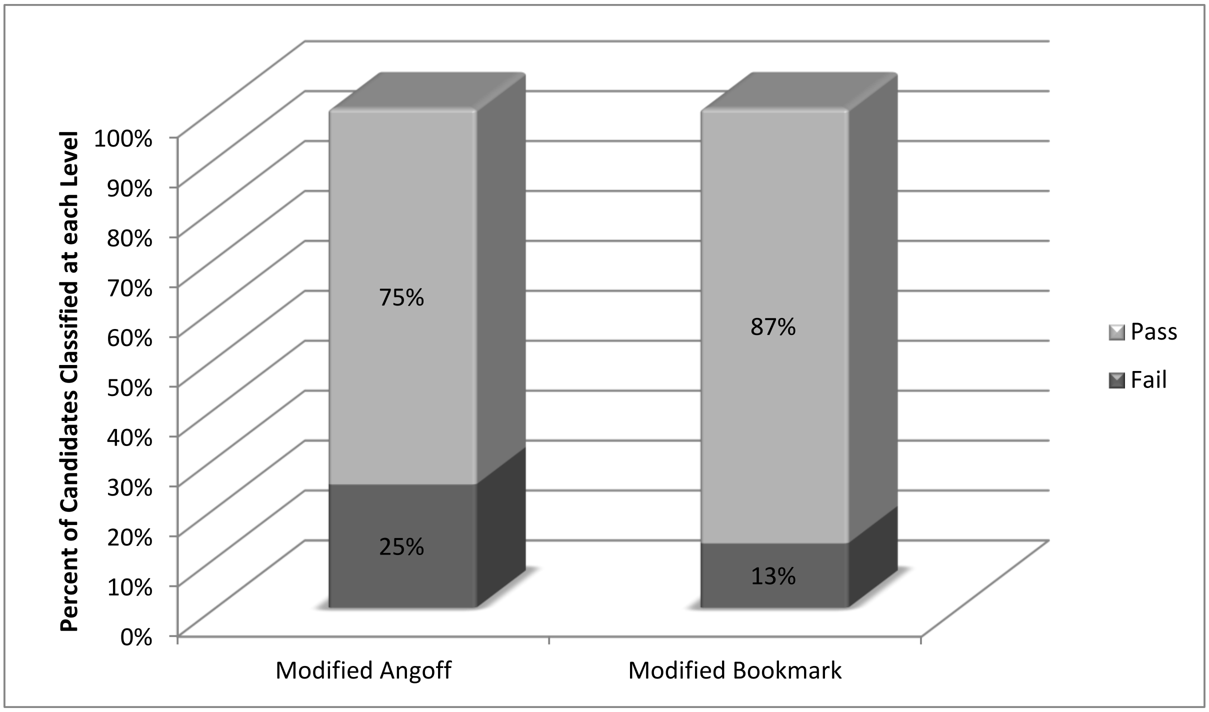

Figure 4. Licensure example – comparison of impact by method

Table 3. Licensure Example - Comparative Analysis of Validity Evidence

Criteria	**Modified Angoff**	**Modified Bookmark**	**Strengths or weaknesses of each method.**
Procedural			
1. Were the panelists qualified to make the judgments that were asked of them?	Yes	Yes	
2. Was the panel representative of the population who should be involved?	Yes	Yes	
3. Did the panelists have a good understanding of the test, the purpose of the test, the intended use of scores, and the abilities of the intended population?	Yes	Yes	
4. Did the panelists have a good understanding of the performance levels for which they were setting cut scores?	Yes	Yes	
5. Were the panelists properly trained on the standard setting method and prepared to make the judgments that were asked of them?	Yes	Yes	
6. Was the standard setting method appropriate for the test, the intended use of scores, and the panelists?	Yes	No?	Given the partial credit nature of the items, the benefits of the Bookmark method may not have been realized because of the understanding of the scoring guide.
7. Was the standard setting method carried out in the way in which it was designed – or were any adjustments justified and appropriate?	Yes	No	Item level scores for the Bookmark method were not based on item response theory values but rather on raw score performance and weighted values assigned to different response options.
8. Was an iterative processes used so that panelists were able to review their recommended cut scores before they were finalized?	Yes	Yes	
9. Were the panelists provided data as input to their decisions (e.g., impact, item difficulty)?	Yes	Yes	
10. Does the evaluation data support the panelists' understanding of the methodology and their confidence in their judgments?	Yes	Yes	
Internal			
11. Do the ratings (recommended cut score and item-level if applicable) suggest the panelists were in general agreement with one another?	Yes	Yes	Smaller standard deviations observed in round 2 data from Bookmark method as compared to Angoff
12. Do the panelists' item-level ratings correlate with available measures of item difficulty? *(for test-based methods)*	Yes	Yes	The Bookmark method requires panelists to provide ratings that follow the estimates of item difficulty. Moderate correlations observed with Angoff results.
13. Do the panelists' examinee-level ratings correlate with other measures of examinee ability? *(for examinee-based methods)*	N/A	N/A	
External			
14. If performance level expectations were developed during the standard setting – are they reasonable and consistent with other	Yes	Yes	

Criteria	Modified Angoff	Modified Bookmark	Strengths or weaknesses of each method.
documented expectations?			
15. If applied, would the recommended cut scores result in a reasonable classification of examinees into performance categories?	Yes	No	Post-study comparison of impact to historical performance and institution-level ratings of candidates suggests closer alignment with Angoff method.

Through the procedural and internal components, there is moderate supporting validity evidence for both methodologies; more so for the modification of the Angoff method in this instance. The only substantial difference observed was within the external component when the respective impact of each methodology was compared to the expectations of the faculty working directly within the professional training programs. This analysis found the results of the Angoff methodology more closely aligned with the expected examinee classifications. Overall, these results suggest that when forced to factor item difficulty into their decisions about what a MQC would likely answer correctly (e.g., Bookmark), the panelists had lower expectations for performance as compared to the Angoff method when they could pick any set of items that they felt the MQC would likely answer correctly. This can happen when panelists disagree with the item ordering as presented in the OIB – panelists tend to error on the side of a lower cut score. There is strong validity evidence for both methodologies. However, because of the concern about the impact of the Bookmark cut score relative to the external criteria (training program performance), a greater emphasis should be placed on the Angoff results when presented to policy makers

Conclusions

In our experience, if policymakers do not fully understand how standard setting data were collected or what it means, they may rely on inappropriate information or intuitive perceptions about cut scores for making decisions (e.g., selecting a value close to 70% or one that approximates their view of a good passing rate). Therefore, part of the practitioner's role in the standard setting process is to fully explain the results of the standard setting process and provide interpretive guidance to disentangle the variability within the results when multiple methods are used.

With this paper we hope to have started a conversation about how practitioners can interpret divergent cut scores that resulted from the use of multiple standard setting methods. As researchers in our field have concluded, there is not a best method, nor is there a true cut score. We will follow this line of thinking and say that when multiple standard setting methods are applied in a given situation, there may not be a best set of results that disqualify other divergent results. The framework outlined in this paper represents much of the guidance provided by others in the field on what is needed to have a defensible standard setting process. Our efforts were targeted at highlighting the evaluative elements that could identify relative strengths and weaknesses of each standard setting process.

When applied systematically, practitioners could use this framework to organize the strengths and weakness of each method as applied to their exam and use this to help policymakers focus their judgments on the appropriate aspects of standard setting evidence. Policymakers can then use this summative recommendation from a practitioner and consider the political and practical elements of setting passing standards (e.g., tolerance for Type I and Type II errors) in the final decision rule.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement,* (2nd ed., pp. 508-600), Washington, DC: American Council on Education

Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 12-21.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G., & Bunch, M. (2007). *A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: CA: Sage Publications.

Geisinger, K., & McCormick, C. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 10*(2), 17-22.

Geisinger, K. (1991). Using standard-setting data to establish cut-off scores. *Educational Measurement: Issues and Practice, 29*(1), 38-44.

Green, D. R., Trimble, C.S., & Lewis, D.M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice, 22*(1), 22-32.

Hambleton, R. K., & Pitoniak. M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger Publishers.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Hofstee, W.K. (1983). The case for compromise in educatiuonal selection and grading. In S. B. Anderson & J.S. Helmick (*Eds.), *On educational testing* (pp. 109-127). San Francisco: Jossey-Bass.

Impara, J. C., & Plake, B. S. (1997). An alternative approach to standard setting. *Journal of Educational Measurement, 34(4)*, 355-368.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: MacMillan Publishing Co.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64(3)*, 425–461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Mitzel, H.C., Lewis, D.M., Patz, P.J., & Green, D. G., (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.

Zieky, M. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives* (pp. 19-52). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.