Are They Mission Ready?
Using the Modified Angoff Method To Set Cut Scores

Ingrid Mellone, Carol Faben
Camber Corporation
Orlando, Florida
imellone@camber.com, cfaben@camber.com

ABSTRACT

Formal assessment is well established in the military and government for applications such as initial selection, promotion, and end-of-course training. For end-of-course assessments of lengthy and/or critical training, it is particularly important for leadership to be confident in the passing score required. Qualified people must not be excluded from passing, and unqualified people should not pass. Yet currently, required passing scores for criterion-referenced tests are often set using arbitrary methods. Although such methods may take into account the criticality of the content overall, they do not use a detailed enough description of job performance requirements to establish "minimally acceptable levels."

This paper describes the importance of establishing a rational passing score, or cut score, and several ways of establishing cut scores, focusing on the Modified Angoff (MA) method. This widely used conjectural method has been adjudicated in the courts and is therefore considered defensible. The MA method features a group of informed judges independently estimating what proportion of minimally qualified test takers will correctly answer each test question. Advantages and disadvantages of the method are discussed, as well as factors in its successful application. The authors have employed the MA method for several years on behalf of the Veterans Benefits Administration (VBA) Skills Certification program, a system of Congressionally mandated, high-stakes certification tests. Although the MA method may be applied to a variety of assessment tests and formats, the VBA tests are comprised of multiple choice and similar test item formats, and are delivered online. The process used to collect judges' estimates is discussed, including the frame-of-reference training provided, the technology supporting the intake of ratings, and the computation of cut scores for these tests. Compared with arbitrary methods, the MA method provides greater assurance that those who pass are, indeed, qualified to pass.

ABOUT THE AUTHORS

Ingrid Mellone, MS, CPT has over 20 years of experience analyzing, designing, developing, implementing, and evaluating technology-based learning and performance solutions for Government and nonprofit customers. She holds a master's in Instructional Design and Development and the Certified Performance Technologist (CPT) designation from the International Society for Performance Improvement (ISPI). She serves as Skills Certification Project Lead, ensuring the quality of all products, developing or modifying processes to improve team efficiency and effectiveness, and applying test development research concepts and technologies.

Carol Faben, MA, PMP is the Program Manager for the VBA Skills Certification testing program. In this role she is responsible for the full program scope and delivery, client relationship management, budgeting, scheduling, and personnel management. Before joining Camber, she spent nearly 20 years with Educational Testing Service (ETS). She is a certified Project Management Professional (PMP) and a trained Six Sigma Black Belt. Carol also has a master's in Teaching.

Are They Mission Ready?
Using the Modified Angoff Method To Set Cut Scores

Ingrid Mellone, Carol Faben
Camber Corporation
Orlando, Florida
imellone@camber.com, cfaben@camber.com

INTRODUCTION

The goal of a test, or assessment, is to provide valid inferences about what the test taker should know or be able to do. Tests are given so that informed judgments of employee readiness, learning, and performance can be made. Formal assessment is well established in the military and government for applications such as initial selection (e.g., American College Test (ACT)), promotion (e.g., advancement exams), and end-of-course training (e.g., Naval Nuclear Power School (NNPS)).

For end-of-course assessments of lengthy and/or critical training, it is particularly important for leadership to be confident in the passing standard, or cut score, required. Qualified people must not be excluded from passing, and unqualified people must not pass. Performance on such assessments establishes a record of accountability that can follow employees throughout their subsequent work experiences.

This paper discusses the importance of establishing a rational cut score and several ways of establishing cut scores for criterion-referenced tests. It elaborates on the Modified Angoff (MA) method, a widely used method that has been adjudicated in the courts. Advantages and limitations of the MA method and several common modifications to it are discussed, as well as factors in its successful application. Details about the MA method used to set the cut scores for the Veterans Benefits Administration (VBA) Skills Certification tests of claims processing personnel are discussed.

ESTABLISHING A RATIONAL CUT SCORE

Depending on the purpose of an assessment and the consequences of passing unqualified test takers or not passing qualified test takers, a defensible cut score may be required. "Defensible" in this sense means the process and logic for determining the cut score are documented in a way that is understandable to and will stand up to the scrutiny of stakeholders (Ricker, 2003). Here are a few examples of such assessments:

- Initial selection for military occupational specialty (MOS), where significant resources will be invested subsequently to develop a service member's specialized skills
- High-stakes certification examinations such as licensure exams, where misclassifying a physician's skill level could have severe consequences for the patient
- Final exams for courses covering highly critical content, such as that for costly or dangerous military weapons systems

No matter the arena in which the test is applied, there can be dire consequences if an employee makes certain errors or does not possess the required level of knowledge, skills, or abilities (KSAs). One's job, status, salary, public perception, and the public itself might suffer. Minimizing this risk is one reason criterion-referenced tests are used. Such tests are designed to assess test takers' mastery of certain KSAs without regard to the performance of others (Shrock & Coscarelli, 2007). In criterion-referenced tests, the employee must demonstrate a defined level of knowledge of the job domain. This is contrasted with norm-referenced tests (e.g., advancement exams), which compare the performance of employees with one another. This paper focuses on criterion-referenced tests.

If it is important to have confidence in the cut scores set for criterion-referenced tests, why are they often set using an arbitrary method? For example, the "70 is passing" scale familiar from high school might be used or the opinion

of one or two decision makers ("Ah, we've used 80 successfully for some time") (Biddle, 1993). Such a method might take into account the criticality of the content overall, but it does not use a detailed enough description of job performance requirements to establish minimally acceptable levels. Finally, professional organizations echo this sentiment by stating that a cut score should be set so as to represent a reasonable expectation of job performance. Cut scores should be determined with respect to professional standards and should not reflect arbitrary percentages, according to the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education, Standards for Educational and Psychological Testing (NCME), 1999).

Kehoe and Olson (2005) describe how to establish a foundation for setting cut scores. This foundation includes defining the purpose of the test, the level of certainty required of the cut score, adjusting the cut score to reflect organizational considerations, and the required level of job performance. The required level of job performance is a value judgment, the authors note, even where standards are well articulated (e.g., "…the probability of a radioactive release should be no greater than 10^{-7}…" (p. 419)). Only after these foundational decisions are defined should the process for setting cut scores commence.

COMMON METHODS FOR DETERMINING CUT SCORES

There are better and worse methods of setting cut scores. Some methods are less subjective than others, more well researched, or more well accepted by the courts. Most involve the cost of convening participants in a central location to be trained and to make their ratings. The process for setting cut scores involves a combination of psychometrics, practical factors, and judgment. It should be noted that subjectivity cannot be eradicated from the activity, although it should be minimized (Kane, 1994; Zieky, Perie, & Livingston, 2008).

This section briefly describes a few of the many standard setting methods used in the assessment field, all of which are more scientific and rigorous than arbitrary methods (see Table 1). These methods fall into two categories, those judging 1) test takers or the products they produce or 2) individual test questions or bodies of questions. Although many methods may be used for both performance and knowledge/skill-based tests, the judgments of questions methods are more commonly used for knowledge/skill-based tests.

Table 1. Common Standard Setting Methods

Method	Judgment of…	Appropriate Usage		Example
		Performance Tests	Knowledge / Skill Tests	
Informed Judgment	People or products	✓	✓	A pharmaceutical firm requires lab technicians to analyze 3 specimens with 100% accuracy.
Contrasting Groups	People or products	✓	✓	A professional organization gives its test to a group of competent facility managers (1) and a group who have not had training (2). The cut score is set beneath group 1's performance and above group 2's.
Bookmark	Test questions		✓	Several states use the bookmark method to set the required passing scores for different achievement levels on K-12 assessments.
Modified Angoff (MA)	Test questions	✓	✓	Firefighters must achieve 80% on a multiple choice test and drag a 170-lb dummy for 100 ft.

Refer to Zieky, Perie, & Livingston (2008) for a complete discussion of these and other methods.

Judgments of People or Products

The following methods are based on the performance of groups of test takers.

Informed Judgment
In this method a group of subject matter experts (SMEs) and those dependent on test takers' satisfying the standards (e.g., test takers' supervisors) are polled to decide the number of criteria candidates must satisfy to pass the test (e.g., 90%, for safety). To do this, the SMEs may evaluate a variety of sample products (e.g., essay, welded joint, dessert) or performances (e.g., patient interaction) in either a testing situation or in the workplace, and possibly over an extended period of time (Hale, 2010; Zieky, Perie, & Livingston, 2008). An advantage of this method is that SMEs are generally able to perform the task of standard setting. A disadvantage is the substantial time it can take to evaluate actual products.

Contrasting Groups
With this method two groups are identified to take the test: a known competent group (masters) and a group known not to possess the KSAs tested (non-masters). Their scores are compared and the cut score is set between the mean scores of the groups. If the organization wants to decrease the chances of passing "false positives"—people who are known to lack the KSAs—it sets the cut score higher, close to one standard deviation (SD) below the mean score of the group known to possess the KSAs. Conversely, if it wants to decrease the chances of failing "false negatives"—people who are known to lack the KSAs—it sets the cut score lower, close to one standard deviation (SD) above the mean score of the group known not to possess the KSAs (Hale, 2010). According to Zieky, Perie, & Livingston (2008), SMEs are readily able to follow prescribed procedures. A criticism is that it may be difficult to identify a large enough sample size of known masters and non-masters (Shrock & Coscarelli, 2007).

Judgments of Questions

The following methods are based on SME predictions of how test takers will perform on individual test questions, or test items. These methods are considered conjectural methods.

Bookmark
The Bookmark method was developed to be used with tests that are scored using item response theory (IRT), which involves applying mathematical models to testing data (Zieky, Perie, & Livingston, 2008). Statisticians first create a booklet of questions appearing on a test, sequenced in order of estimated difficulty. Judges next identify the point in the booklet (place a bookmark) where the borderline test taker would probably answer correctly (i.e., have a .67 probability of answering correctly). The bookmark is then translated into the cut score using a specialized software application. This method is increasingly used for large audiences, in the K-12 sector, and where multiple cut scores are needed (e.g., to identify basic and advanced skill levels). One criticism is the advance effort required to prepare the ordered item booklets. In addition, although judges are able to place the bookmark, they might be unfamiliar with IRT and its mathematical concepts, and therefore might not understand how the bookmarking activity is transformed into the cut score.

Modified Angoff (MA)
The initial description of the Angoff method appeared as a footnote in a lengthy chapter that William H. Angoff authored in Thorndike's *Educational Measurement* (1971): "…judges would think of a number of minimally acceptable persons…and would estimate the proportion…who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score." The minimally acceptable person is commonly referred to as a "just sufficiently qualified," or JSQ, test taker. The job performance of JSQs just *barely* meets the minimum standard of performance. There may be certain content areas that JSQs need only be familiar with, some they should have working knowledge of, and some they should have in-depth knowledge of. A group of informed judges independently estimate what proportion of JSQ test takers will correctly answer each question. A probability of 95% means that almost all qualified test takers will answer correctly. A probability of 25%, for a four-response multiple choice question, means the question is very difficult and only one-quarter are estimated to answer correctly. The ratings of the judges are then averaged to yield the cut score for a given test question. Finally, the cut score is adjusted downward one, two, or three standard errors of measurement (SEM) to account for measurement error in the test (Biddle, 2006). This adjustment comprises the Modified Angoff (MA) method. The decision of how many SEMs to adjust is based on statistical and human factors, and provides a

confidence interval around the cut score (i.e., a decreased risk of passing the unqualified and failing the qualified). Advantages and disadvantages are discussed in the next section.

ELABORATING ON THE MODIFIED ANGOFF (MA) METHOD

Advantages

Cut score setting methods have been evaluated on several criteria, including expert opinion, legal decisions, technical sufficiency, and practical considerations as described by professional standards and recommendations (*Standards for Educational and Psychological Testing* (AERA, APA, NCME; 1999), *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003), and *Uniform Guidelines on Employee Selection Procedures* (U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978)). The MA method is widely accepted because it performs well on these points, because it is relatively simple to administer, and because it requires only minimal training to implement. In addition, depending on the particular implementation design, the method can be an efficient use of SME time. According to Biddle (2006), 4 to12 is an adequate number of SMEs.

The MA method has been adjudicated in the courts, finding statutory basis in Title VII of the 1964 Civil Rights Act as amended by the Civil Rights Act of 1991. For example, in *Lanning v. Southeastern Pennsylvania Transportation Authority (SEPTA)* (1999), the Court said that a cut score should correspond to the "minimum qualifications necessary to perform successfully." *Gulino v. Board of Education* (2003) similarly upheld a challenged test where the cut scores recommended by experts were estimated to represent the "minimum level of knowledge and academic skills necessary for competent performance…." Some organizations (e.g., the military) may not anticipate a legal challenge to their tests, but application of the just sufficiently qualified (JSQ) definition in the cut score activity is a valuable component of the MA method. Recognition of the value of the JSQ's detailed description of required job performance in large part accounts for the MA method's popularity.

Constraints

Judgmental cut score methods in general and the MA method in particular are criticized for their assumption that a group of SMEs can apply a definition of the JSQ test taker in their ratings. Some researchers (e.g., Berk, 1996) have said the cognitive load required is too high. To address the concern, it is critical that qualified SMEs be selected. SMEs should be knowledgeable of the relevant content and the community of test candidates. It is also important to train the SMEs to provide their judgments. A further limitation is that the resulting cut score "…does not specify which, if any, items *must* [emphasis added] be answered correctly to reach that cut score" (Ricker, 2003, p. 14). A strategy for overcoming this limitation is to calculate separate cut scores for each dimension (category) on a test.

Modifications

Even though the MA method is very popular, numerous modifications to it have been devised in an effort to simplify the process or to increase the reliability of results. Several of these modifications are described below. The research literature shows that practitioners often include multiple modifications in the cut score process. As shall be seen, some modifications are more supported than others by the literature or require more skill in execution.

1. Hold a Discussion Period
Using an iterative feedback and rating process wherein judges have an opportunity to discuss their initial ratings has been found to decrease variability among ratings (Busch & Jaeger, 1990). Two or more discussion rounds may be spaced over hours or weeks. Busch and Jaeger caution that during the discussion period the panel must be protected from the undue influence of strongly opinionated judges.

2. Provide a JSQ Definition
Providing judges a definition of minimally qualified personnel ensures that judges have a shared mental model of the JSQ candidate while they are making their ratings (Fehrmann, Woehr, & Arthur, 1991). Like modification 1, modification 2 has been found to decrease variability among ratings.

3. Provide Performance Data
Some practitioners provide judges the percentage of test takers answering each item correctly (*p*-values) before finalizing their ratings (Busch & Jaeger, 1990). This is referred to as a reality check. This modification is controversial because although providing data leads to decreased variability among judges' ratings, it could undermine the logic of the Angoff method. If judges depend too heavily on performance data and simply adjust their ratings to reflect that data, the judges could equate average performance with required performance (Clauser, Mee, Baldwin, Margolis, & Dillon, 2009).

4. Make a "Yes/No" Decision
With this simplification of the MA method, judges decide whether a single JSQ candidate would (Yes) or would not (No) correctly answer a given item. This modification has been shown to produce similar cut scores compared with other methods (Impara & Plake, 1997). However, when combined with other modifications, researchers found mixed results (Chinn & Hertz, 2002).

5. Apply Item Response Theory (IRT)
Rasch IRT models have been used to calculate item difficulty instead of collecting judges' ratings or as a means of evaluating judges' ratings (Ricker, 2003; Hsieh, 2013). As with the Bookmark method discussed above, using IRT with the MA method requires complicated calculations, specialized software, and large sample sizes.

6. Other Modifications
Although the MA method is popular and mature, researchers have continued to compare various combinations of modifications in an effort to identify best practices and efficiencies.

- Jalili, Hejri, & Norcini (2011) compared the MA method with a three-level Angoff (TLA) method in setting the cut score for a clinical exam for medical students. TLA is similar to modification 4 above, but in addition to Yes and No, allows judges to provide a third "Maybe" rating. Two TLA methods were assessed: TLA with a discussion period (modification 1 above) and TLA with reality check using performance data (modification 3 above). In both TLA methods judges retained independence in their ratings. Final ratings for each station were averaged. Ratings for all stations were averaged again to yield the standard for the overall test. The different methods resulted in considerably different cut scores, with the MA with the reality check providing more stable cut scores. Stable cut scores are desirable because they imply reliability in the standard that is applied when the test is given multiple times.
- Hoffman, Tashina, and Luck (2010) report on a variant of modification 3 above in a law enforcement exam used for promotional purposes. Judges were provided 9 items with *p*-values ranging from .97 to .20 as a "difficulty anchored" rating scale to use as a guide while making ratings of items included on the test. The particular items included on the scale were from a previous test given to the same candidate pool but did not appear on the test for which cut scores were being determined. Therefore, they did not comprise performance data in the strict sense. After a brief training period, SMEs independently rated all items with no discussion. The researchers argued that their method struck a reasonable compromise between providing no or complete normative data on items. They argued that the limited data that was provided did not remove the judgmental aspect of the rating process. The authors found considerable support for their use of the scale, citing reliability and correlation data.

APPLYING THE MODIFIED ANGOFF METHOD IN VBA SKILLS CERTIFICATION

The precise cut score workshop design to employ for a given test depends on the purpose of and consequences for the test as well as other factors. VBA's Skills Certification Program is a system of criterion-referenced assessments of individual job knowledge/skill. The Program is comprised of over half a dozen tests for technical and managerial positions. VBA began developing the Skills Certification Program in 2003 to improve the organizational performance and professionalism of claims processing personnel and in response to a mandate from Congress. When VBA employees pass a test they become certified. They have demonstrated they are qualified to perform the essential duties of their job occupation. Before a test for a position is developed, the requisite duties are documented through a job task analysis.

VBA documents the test policies governing the tests via Memoranda of Understanding it has negotiated with the employee unions. A Design Team for each test serves as VBA's governance committee throughout test

development activities. Teams of experienced SMEs write and review the items for each test in regularly scheduled, facilitated workshops. Each test is typically offered online twice a year at as many as 57 VBA regional offices in the continental U.S., Puerto Rico, Hawaii, and the Philippines. VBA uses the results of testing to support employee development, to provide insight into workforce capability, and as input into the strategic planning of training.

In the cut score workshops convened for VBA, modifications 1 and 2 are used as well as *p*-values for the practice items (after Hoffman, Tashina, and Luck, 2010). We selected the MA method and these modifications because they are a good fit for the purpose and consequences of the VBA tests and for technical and practical reasons. Details of the MA method used for VBA are described below.

Preparing To Set Cut Scores

The first step in determining the cut score for a VBA test is for the Design Team to define the certification-level performance standard (i.e., the minimally qualified, or JSQ candidate). As this definition is used in multiple applications for a test, the Design Team reviews the JSQ definition regularly to ensure it continues to align with VBA expectations. This occurs before a cut score workshop is convened.

VBA then identifies 7-10 representative subject matter experts (SMEs) who are very familiar with the target position to act as judges in the cut score workshop. SMEs include incumbents, supervisors, trainers, and other stakeholders such as union representatives. Through the MA method described next, participants apply the JSQ definition the Design Team has developed.

Training the SMEs

At the beginning of a workshop, the facilitator explains that through the workshop the SMEs will determine the cut score for the test items—and, therefore, the test. The facilitator then conducts frame of reference (FOR) training (Fehrmann, Woehr, & Arthur, 1991). FOR training is designed to instill a common conceptualization of required test performance for eligible candidates (i.e., the JSQ definition). The primary purpose of this activity is to generate a necessary and sufficient level of consensus and consistency in the ratings the judges provide. The group receives the previously written JSQ definition and discusses the conceptual JSQ test candidate in depth, including the level of knowledge JSQ candidates should have and the performance levels they should exhibit on the job for the various content areas of the test (modification 2 above). Participants are then able to review each item in comparison to the JSQ definition and develop a cut score for it.

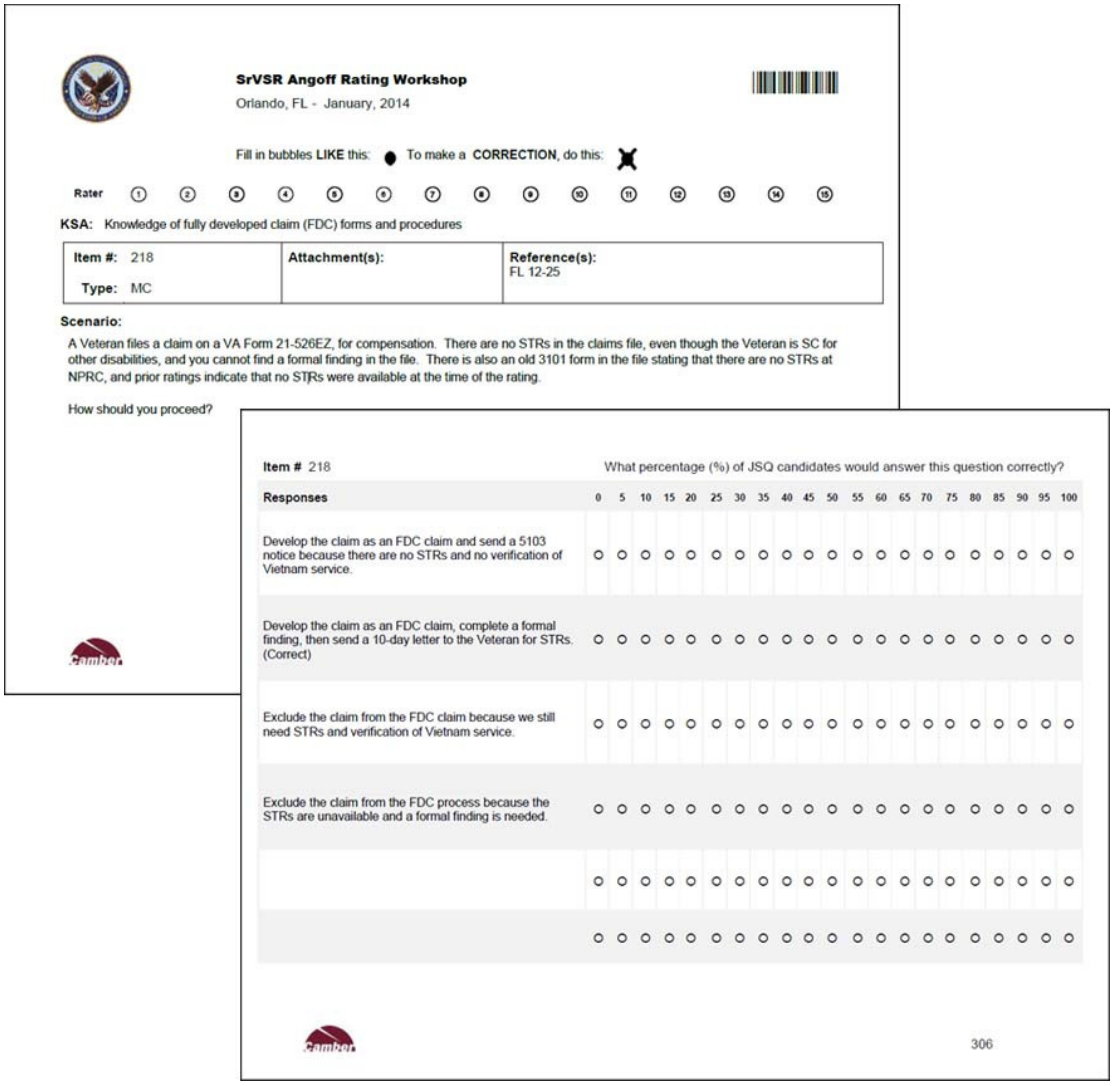
SMEs are next given a small set of practice items to rate. Item types include multiple choice, fill-in-the-blank, and situational judgment test (SJT) (Gunter, Mellone, Oakley, & Faben, 2013). SMEs are instructed to make independent ratings on each of the test items, answering the question, “What percentage of JSQ candidates would answer this question correctly?” by filling in a bubble on a scannable answer sheet (see Figure 1).

A facilitated discussion period follows in which each participant discusses his/her ratings and reasons for the ratings. Discussion may revolve around the complexity of the content, the clarity of the regulations for performing a procedure, the nature of the training provided on the topic, or the frequency with which errors are encountered related to the item. The SMEs may then make adjustments to their ratings in response to what they have heard (modification 1 above). In addition, in a variant of the difficulty anchored rating scale discussed above (Hoffman, Tashina, and Luck, 2010), *p*-values and existing cut scores for the practice items are shared as feedback to the SMEs. In providing this feedback the facilitator is careful to distinguish the standard that has been set for these items (the cut score) and what *p*-values represent (the *average* performance of test takers).

Assigning Cut Scores

Unless additional practice is needed, participants are next assigned test items in batches of approximately 15 to rate on their own. The initial individual ratings for each batch are then reviewed in roundtable fashion. The facilitator informally records each rating. If ratings differ significantly (e.g., by more than 20 percentage points), the SMEs are given an opportunity to discuss their rationale for assigning a rating. Participants may change their ratings based on the discussion if desired, but changes are not required. Final SME ratings are then collected. Throughout the workshop the facilitator encourages participants to consider what the SMEs share during the discussion period,

especially if they tend to provide systematically high or low ratings. In addition, SMEs are regularly reminded of the JSQ definition they should be using as the basis for their ratings.



SrVSR Angoff Rating Workshop
Orlando, FL - January, 2014

Fill in bubbles **LIKE** this: ● To make a **CORRECTION**, do this: ✖

Rater ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬ ⑭ ⑮

KSA: Knowledge of fully developed claim (FDC) forms and procedures

Item #: 218 **Type:** MC	**Attachment(s):**	**Reference(s):** FL 12-25

Scenario:

A Veteran files a claim on a VA Form 21-526EZ, for compensation. There are no STRs in the claims file, even though the Veteran is SC for other disabilities, and you cannot find a formal finding in the file. There is also an old 3101 form in the file stating that there are no STRs at NPRC, and prior ratings indicate that no STRs were available at the time of the rating.

How should you proceed?

Item # 218 — What percentage (%) of JSQ candidates would answer this question correctly?

Responses	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
Develop the claim as an FDC claim and send a 5103 notice because there are no STRs and no verification of Vietnam service.	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Develop the claim as an FDC claim, complete a formal finding, then send a 10-day letter to the Veteran for STRs. (Correct)	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Exclude the claim from the FDC claim because we still need STRs and verification of Vietnam service.	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Exclude the claim from the FDC process because the STRs are unavailable and a formal finding is needed.	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

306

Figure 1. Sample Cut Score Worksheet, Front and Back

Processing Cut Scores

The sheets containing the cut score for each item are scanned; and the results are analyzed and incorporated into the overall test score. Table 2 provides a notional example of these calculations using a 20-item test where each item is worth 1 point. Note: The actual VBA tests have far more than 20 items.

The final ratings of the SMEs are averaged to determine the cut score for each item, as averaging represents their consensus opinion on that item. To account for any measurement error in the test, the cut score is then adjusted downward by two SEMs (Biddle, 2006). Finally, the cut scores for each item are summed and rounded to develop the cut score for a given test. In the Table 2 example, the cut score is 15 out of a possible 20.

Table 2. Example of MA Ratings Translated into Final Cut Score

Test Item No.	Average MA Rating	SEM	MA Rating - 2 SEM
1	0.783	0.025	0.733
2	0.870	0.015	0.840
3	0.707	0.032	0.643
4	0.867	0.017	0.833
5	0.861	0.014	0.833
6	0.633	0.031	0.571
7	0.861	0.022	0.817
8	0.764	0.026	0.712
9	0.779	0.036	0.707
10	0.856	0.019	0.818
11	0.615	0.029	0.557
12	0.856	0.013	0.830
13	0.758	0.022	0.714
14	0.850	0.012	0.826
15	0.643	0.038	0.567
16	0.850	0.012	0.826
17	0.704	0.023	0.658
18	0.733	0.031	0.671
19	0.615	0.023	0.569
20	0.850	0.019	0.812
Test Cut Score			**14.801, rounded to 15**

Reflections on Results

Participant comments received do not support Berk's (1996) assertion that the standard setting task is too cognitively difficult. Here are a few recent examples of satisfaction survey results taken in at the end of the workshop:

- The prep for Cut Score was very effective at preparing to actually do it.
- Training prior to beginning the process was very helpful.
- Great "calibration" among team members.

Through their ratings, participants are able to effectively contribute their knowledge of the position and their community and make the needed judgments. This may be because of the particular MA modifications applied, the format of the worksheet, the use of experienced facilitators, the particular SMEs selected, or other factors.

Because of the applied nature of this work, initial ratings are not retained. Thus, no analysis can be done of initial ratings compared with final ratings collected after the discussion period. As a result, we cannot refute or support Busch & Jaeger's (1990) findings that discussion decreases variability among judges' ratings. Anecdotally, however, SMEs' independent ratings tend to be increasingly aligned as a workshop proceeds and more groups of items are rated. The longer a workshop goes, the fewer ratings have a greater than 20 percentage points difference across SMEs. This suggests that SMEs' mental model of the JSQ is indeed shared. This observation agrees with Hambleton's (1999) observation that the impact of the feedback and discussion period (modification 1) is consensus among participants.

As a means of reviewing a sample of SME judgments, a comparison of observed test difficulties (*p*-values) and adjusted item cut scores for two recent 100-item tests is shown in Figure 2. The calculations plotted are the *p*-value minus the adjusted cut score for each test item.

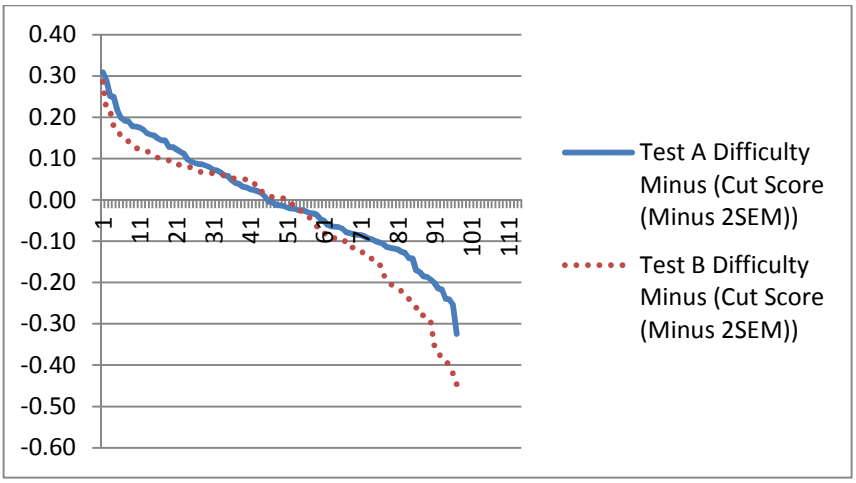


Figure 2. Comparison of Observed Test Difficulties and Adjusted Item Cut Scores

Positive values on the y axis reflect that the SMEs set a standard for the item that was below the average score for the item (the SMEs thought the item would be more difficult for the JSQ candidate than it was). Negative values on the y axis reflect that the SMEs set a standard for the item that was above the average score for the item (the SMEs thought the item would be less difficult for the JSQ candidate than it was). There is approximately the same number of items overestimating as underestimating how the average performer would do. The middle of the graph demonstrates that both tests have a number of items where there is little difference between estimated and actual average scores. In addition, the flatter curve for Test B reflects slightly less variability than Test A, i.e., cut scores diverge comparatively less from average item difficulties. The authors make no conclusions about this data. If participant ratings and difficulty values were the same, all that would reveal is that SMEs were able to estimate the average difficulty of the item. Agreement says nothing about the performance standard required, because average is not necessarily "minimally qualified." Additional study of other tests is planned and will provide data to support a continuous improvement effort. Longitudinal trends within tests and other studies may be performed.

CONCLUSION

Compared with arbitrary methods, the MA method provides much more substantial assurance that those who pass are, indeed, qualified to pass. In an increasingly tight fiscal environment, such assurances can be translated into training dollars saved. Despite this assurance, there is no "true" cut score that is somehow separate from organizational culture. SME cut score participants are making judgments based on their knowledge of the content and their community. The MA method provides a systematic, detailed process that enables SMEs to determine the cut score for a criterion-referenced test. Importantly, this method takes into account item-level difficulty in relation to the expressed standard expected of minimally qualified personnel. Arbitrary cut scores are indefensible, but MA methods produce data that supports further evaluation and refinement of the assessment process. The MA method has been scrutinized by the courts, it meets published standards for assessment, and it is relatively simple to perform. The MA method should be used more widely when setting or resetting the required passing score for tests of lengthy or critical training.

ACKNOWLEDGEMENTS

The authors wish to express their appreciation to the Veterans Benefits Administration for enabling them to share these findings with other organizations and to Dr. Stephen Gunter for his contributions to establishing the MA method employed.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) Joint Committee. (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.

Berk, R. A. (1996). Standard Setting: The Next Generation (Where Few Psychometricians Have Gone Before!). *Applied Measurement in Education, 9*, 215-235.

Biddle, D. (2006). *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing* (2nd ed.). Burlington, VT: Ashgate.

Biddle, R. E. (1993). How To Set Cutoff Scores for Knowledge Tests Used in Promotion, Training, Certification, and Licensing. *Public Personnel Management, 22(1)*, pages unknown.

Busch, J. C., & Jaeger, R. M. (1990). Influence of Type of Judge, Normative Information, and Discussion on Standards Recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*, 145-163.

Chinn, R. N., & Hertz, N. R. (2002). Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education*, *15*, 1-14.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' Use of Examinee Performance Data in an Angoff Standard-setting Exercise for a Medical Licensing Examination: An Experimental study. *Journal of Educational Measurement, 46*(4), 390–407.

Fehrmann, M. L., Woehr, D. J., & Arthur, W. Jr. (1991). The Angoff Cutoff Score Method: The Impact of Frame-of-reference Rater Training. *Educational and Psychological Measurement, 51*, 857-872.

Gulino v. Board of Education, 96 Civ. 8414 (S.D. N.Y., 2003).

Gunter, S., Mellone, I., Oakley, K., & Faben, C. (2013, Dec). Evaluate Training and Performance Effectively, Quickly, and Inexpensively Using the Situational Judgment Test (SJT). Paper presented at I/ITSEC annual meeting, Orlando, FL.

Hale, J. (2010). Performance-based Certification: How To Design a Valid, Defensible, Cost-effective Program. San Francisco, CA: Pfeiffer.

Hambleton, R. K. (1999). Setting Performance Standards on Assessments and Criteria for Evaluating the Process. Presentation at the Edward F. Reidy Interactive Lecture Series, RILS 1999 conference, sponsored by Center for Assessment, October 14-15, 1999, Providence, RI.

Hoffman, C. C., Tashima, C. C., & Luck, G. (2010). Using a Difficulty-Anchored Rating Scale in Performing Angoff Ratings. *International Journal of Selection and Assessment, 18(4)*, 407-416.

Hsieh, M. (2013). An Application of Multifaceted Rasch measurement in the Yes/No Angoff Standard Setting Procedure. *Language Testing, 30*(4), 491-512.

Impara, J. C., & Plake, B. S. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement, 34*, 353-366.

Kane, M. T. (1994). Validating the Performance Standards Associated with Passing Scores. *Review of Educational Research, 64*, 425-461.

Kehoe, J. F., & Olson, A. (2005). Cut Scores and Employment Discrimination Litigation. In F. J. Landy (Ed.), Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives (pp. 410-448. San Francisco: Jossey-Bass.

Lanning v. Southeastern Pennsylvania Transportation Authority, 181 F.3d 478 (3d Cir. 1999).

Ricker, K. L. (2006). Setting Cut Scores: Critical Review of Angoff and modified-Angoff Methods. *Alberta Journal of Educational Research, 52(1)*, 53-64.

Shrock, S. A. & Coscarelli, W. C. (2007). *Criterion-referenced Test Development: Technical and Legal Guidelines for Corporate Training*. San Francisco, CA: Pfeiffer.

Society for Industrial and Organizational Psychology, Inc. (2003, 4th ed.). *Principles for the validation and use of personnel selection procedures.* College Park, MD.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: Educational Testing Service.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978). *Uniform Guidelines on Employee Selection Procedures*. Federal Register, Volume 43, Number 166, 38290-38315.