

Setting Cut Scores:

Critical Review of Angoff and Modified-Angoff Methods

Kathryn L. Ricker

Centre for Research in Applied Measurement and Evaluation

University of Alberta

Edmonton, Alberta, Canada

### Abstract

This paper presents a critical review of the Angoff (1971) and Angoff derived methods, according to criteria for assessing cut score setting methods originally proposed by Berk (1986) and further recommendations by Hambleton (2001). The criteria have been updated to reflect the progress that has been made in standard setting research over the past 17 years. The paper also discusses the assumptions of the Angoff method, and other current issues surrounding this method. Recommendations for using the Angoff method are made.

## Standards for Cut Scores:

## Critical Review of Angoff and Modified-Angoff Methods

## Introduction

There has been a decided political movement towards standardized tests in North America. In Canada, all provinces, with the exception of Saskatchewan and Prince Edward Island, use some sort of standardized provincial examination. Additionally, all provinces participate in the national Pan-Canadian (formerly SAIP) test, as well as international tests including the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Studies (TIMSS).

Without question, the largest scale impetus towards standardized testing can be witnessed in the United States, with the advent of the Public Law 107-110, *No Child Left Behind Act* of 2001 (United States Government, 2002). The act includes assessment and measurement of student progress as one of four main pillars. As a result, each state is federally mandated to develop state-wide tests of student achievement in core (mathematics and reading currently, science by 2005-2006) areas of school curriculum in grades 3 to 8. Over \$770 million in funding has been allocated to aid states in the development and administration of these tests for 2002-2003 alone (United States Government).

One of the purposes of these standardized tests is to increase “accountability” among educators and students. As such, students are expected to meet some standard of proficiency that the tests are designed to assess. Ideally, this standard will be the embodiment of the learning objectives. The standard should represent “mastery” of the learning objectives, or some level of basic proficiency necessary to move on to the next

level, or to function in the real world (van der Linden, 1982). In effect, establishing a standard can be conceptualized as policy making, that has an impact on everyone involved in the testing procedure (Kane, 2001).

Once established, a standard is translated into a cutoff score in the distribution of scores obtained from a set of test items relevant to and representative of the standard. The purpose of the cut off score is to separate examinees who meet the standard from those who do not. But how are these cut scores determined?

The distinction between standard and cut score setting is often a confusing one, since generally one task is not done without the other. However, it is important to keep in mind that standard setting is a philosophical and policy making activity, while setting a cut score is the operationalization of that policy. While the Angoff and other methods are often referred to as methods for standard setting, they are actually used once standards, or a useful form of the standards, have been established.

Given that the stakes for educational testing have never been higher, it is critical to examine the methods that are used for cut score setting in educational tests. Which methods are most appropriate (or perhaps least inappropriate) in a given testing situation? Which method or methods provide the most fair, accurate and reasonable separation of students in terms of performance in established standards?

In a footnote to a book chapter, Angoff (1971) inadvertently introduced a method for standard setting that is, using the amount of attention devoted to it in the research context as an indicator, one of the most commonly used method of setting standards today. The original method has been modified in different ways by researchers (e.g., Hambleton & Plake, 1995; Impara & Plake, 1997; Taube, 1997) in an attempt to improve

it. In 1986, Berk published a “consumer’s guide” to standard setting techniques, which included a set of criteria to be used to assess standard setting methods. He also the assessed various cut score setting procedures, including five Angoff-type methods.

This paper will use criteria based on the recommendations of Berk (1986) and Hambleton (2001) to assess Angoff and Angoff-derived methods of standard setting. In addition, this paper will synthesize some of the empirical investigations of Angoff-type methods conducted since 1986 and discuss some of the current debates and issues relevant to Angoff methods. Finally, recommendations for the modification of the Angoff method that best meet the Berk and Hambleton criteria will be made.

### The Angoff Methods

The variants in Angoff methods can be classified as item-judgment methods. Each item on a test is assessed in terms of how likely minimally acceptable or competent candidates (those who would barely meet mastery standards) are to answer that item correctly.

#### *Basic Angoff Method*

The Angoff method, in its most basic form, is seemingly a very simple process. Perhaps its simplicity should not be surprising, given that it arose from footnote in a book chapter (Angoff, 1971, p.515). A group of judges are each asked to (independently) think of a group of minimally competent candidates who would border on the mastery/non-mastery cut-off. The most typical instruction is for judges to think of a pool of 100 candidates who would “just barely” meet the performance criteria. When Angoff first proposed the method, his instruction was to think of only one candidate. However, with the exception of Impara and Plake (1997), the hypothetical pool of candidates is used.

The judges, working independently, then estimate what proportion of that sample of minimally acceptable candidates would answer each item in the test correctly. These  $p$ -values are summed and usually denoted as the Minimum Passing Level for judge ( $MPL_j$ ). The  $MPL$  represents an individual judge's cut score for the test. The mean of these cut scores is the final cut score for the test. The standard error can also be calculated for the cut score. A lower standard error is desirable since it denotes better agreement among the judges (and less uncertainty about where the "true" cut score should lie).

This method does not just apply to minimally competent candidates, but could also be used to create a cut score for any grouping within the population. For example, Angoff methods could be used to set a cut score for a standard of excellence on a test. In this case, judges would be required to conceptualize a group of 'minimally excellent' examinees.

#### *Modifications of the Basic Angoff Method*

Many modifications and interpretations have been made to the Angoff method in an attempt to improve the process, particularly to improve agreement among judges. To add to the confusion, these modifications have been grouped under a common title, "modified Angoff" procedures, often without describing exactly what specific modifications to the original method are used. Presented below is a discussion of the effects of five specific modifications that have been empirically tested and reported in the scholarly literature.

##### *Using an iterative process.*

Conducting a number of iterations of the item assessment process is the most commonly used modification of the Angoff procedure (e.g., Busch & Jaeger, 1990;

Woehr, Arthur, & Fehrmann, 1991; Hambleton & Plake, 1995). Repeatedly setting cut scores is listed as a desirable characteristic of a judgmental process (Hambleton, 2001). The time between rounds is used for discussion among the panel members. The intent of the discussion is to increase the agreement among judges (i.e., to reduce the standard error). Use of two (e.g., Chang, 1999) or three (e.g., Busch & Jaeger, 1990) iterations have been reported in the literature. Busch and Jaeger (1990) found that an iterative process reduced variability among judges for item estimates and the cut scores overall. However, the true value of iterations for cut score setting is limited by the ability of the members of the panel to make independent judgments (Busch & Jaeger, 1990), and of the panel moderator to maintain a process whereby strongly opinionated judges do not have undue influence.

*Presentation of normative data.*

There is some controversy over whether or not it is appropriate to present performance data to judges while using the Angoff procedure. Presentation of normative data has been shown to improve inter-judge reliability. These data are presented to the judges prior to the final iteration of estimating item  $p$ -values. Busch and Jaeger (1990) found that presentation of examinee results to the judges panel caused an increase in the correlation between item  $p$ -values and item difficulties, which in turn led to greater agreement among the judges (i.e., increased inter-judge reliability). Inter-judge reliability is often used as a measure of the quality of the judges' ratings.

How strong is the influence of presenting data to the judges? Norcini, Shea, and Kanya (1988) found that judges used normative data about 25% of the time, and that the average change per item was relatively small, with a tendency for the change to occur on

items that had originally been estimated to have very high or low initial  $p$ -values (suggesting that judges' had initially over- or underestimated the difficulty of an item).

*Yes/no estimation procedure.*

The central concept of the Angoff procedure is *estimating the proportion of borderline examinees* that would correctly answer a test item. To simplify the process, Impara and Plake (1997) proposed that judges decide whether a *single* minimally competent candidate *would or would not answer* an item correctly. Known as the yes/no procedure, the rationales for this modification are that it is easier for a judge to think of a single person than a pool of candidates and to make a simple yes/no decision.

The results of empirical investigation of this modification are equivocal. Impara and Plake (1997) compared the traditional percentage estimation procedure with the yes/no method. They reported that the methods produced essentially the same cut score and that the iterations produced only “fine tuning” of the estimates. They postulated that while individual judges may not be very accurate at predicting individual performance, the judges as a group will be able to produce a reasonable cut score when aggregated. However, Impara and Plake also concede that their results may have been contaminated by the panel conducting traditional method immediately prior to trying this method in one of their experiments.

Chinn and Hertz (2002) compared these same methods in two experiments. Judges reported that the yes/no procedure was much easier to conduct, but their final cut scores were far more influenced by the empirical data that they were presented, and their overall cut scores were less stable than the percentage method over iterations of the process.

*Applying relative weights to scores.*

Hambleton and Plake (1995) developed what they referred to as the “extended Angoff” procedure. They applied Angoff-type procedures to multi-dimensional, polytomously scored items. In addition to predicting scores of borderline candidates on each dimension, raters were asked to provide a weight of each dimension and each exercise, where the weights reflected the relative importance of the content covered by the examination. The scores and weights for each exercise were multiplied together and then summed across exercises to calculate a final cut score for the entire assessment. When evaluating this method, the judges expressed high confidence in the cut score they produced and rated the overall process as very successful. Judges were more confident in their ratings, with less variability in confidence among judges than when the dimensions were not weighted.

*Using item response theory to calculate the final cut score.*

When judges are asked to assess the probability of candidates correctly answering an item, they are in essence determining the difficulty of the item. In effect, the Angoff rating estimates the ability level denoted as  $\theta$  in Item Response Theory (IRT) of a minimally acceptable examinee (Kane, 1987). Taube (1997) extended this idea by using judge’s ratings to work backwards to calculate  $b$ - (difficulty) parameters for each item using a Rasch IRT model, given by:

$$P(\theta)_i = 1 / (1 + \exp(-D(\theta - b_i)))$$

where  $P(\theta)_i$  is the probability of an examinee with a given  $\theta$  correctly answering item  $i$ , and  $D$  is a scaling constant equal to 1.7. Instead of calculating the sum of the item probabilities as the cut score, the mean item difficulty was calculated. Estimated item

difficulties were highly correlated with actual item difficulties, which adds to the validity evidence of the standard setting process.

### Strengths and Weaknesses of Angoff and Angoff Modifications

#### *Presentation of Normative Data*

While introducing empirical data might be a good way to improve inter-judge consistency, comparing the estimated item performance to actual item performance after the judges have been presented these data seems to be somewhat of a self-fulfilling prophecy. Further, the assumption that introducing test results or other impact data is helpful only holds true if examinee performance was not influenced by systematic flaws in the testing procedure (i.e., the examinees performed as expected, and not systematically better or poorer than they might otherwise).

It would seem imperative that judges be expert in the construct of the test and familiar with the population of examinees. The more confident they are about their expertise, the less likely they are to be unduly swayed by empirical data. For example one might expect expert teachers to be less influenced by data than parents or outside stakeholders.

Busch and Jaeger (1990) suggested that presenting impact data is necessary so that test standards are reasonable in relation to the behaviour of the population on the test, in spite of the fact that establishing performance standards and correspondent cut scores implies that judging be impartial and separate from the data. Ideally, normative data would not be introduced into what is supposed to be a criterion-referenced exercise. However, because judging is never perfect, it is necessary for test standards and cut scores to be realistic in order for them to be accepted by stakeholders.

As a final thought on presenting impact data, it is not always clear whether overall group data, or just the borderline group data should be presented, or which, if presented, would be more helpful/useful. This is a question that requires further systematic investigation.

### *The Complexity of the Method*

#### *Cognitive complexity.*

The yes/ no procedure is an interesting idea that warrants further research. Impara and Plake (1997) introduce this idea to simplify the decision making process that judges must undergo for each item. However, there is some evidence to suggest that this modification produces less stable estimates and that judges using this method are more strongly influenced by the presentation of empirical data (Chinn & Hertz, 2002). The question of the validity of this modification area emphasizes the need for cut score setting methods that best balance the need to simplify the procedures for the sake of judges, while still producing a valid and defensible result. If this method is not comparable to the traditional  $p$ -value estimation procedure, then perhaps it would be better to increase judge training and practice, rather than trying to make the method simpler in this way.

#### *Technical complexity.*

There are several drawbacks to using IRT parameter estimates for setting cut scores. First, calculating parameter estimates for an IRT model is a complicated process and requires a specialized computer program. Second, items that do not fit the selected IRT model must be discarded, since parameter estimates cannot be calculated for them. Third, large sample sizes (a minimum of 500 to 1500 examinees, depending on the IRT model being used) are required to establish stable parameter estimates for items. While

the idea has merit, this procedure does not seem practical for most cut score setting applications. However, IRT methods should not be abandoned completely. While not currently feasible in most settings, IRT is becoming increasingly accessible due to improvements in IRT software packages that can calculate parameter estimates.

### *Assumptions of Angoff methods*

#### *The concept of minimal competence.*

One of the biggest assumptions of the Angoff method(s) is that a pool of expert judges will be able to conceive a comprehensive and appropriate picture of what minimally acceptable candidates will “look” like, in terms of their performance on a test (Impara & Plake, 1997). This assumption is the root of much criticism of this method, as well as other judgmental methods. Berk (1996) characterizes conceptualizing a group of borderline examinees as a "...nearly impossible cognitive task..." (p. 216), while Shepard (1995, cited in Plake & Impara, 2001) claimed the process exceeds human cognitive processing capacities. Empirical examinations of this assumption have found mixed results. Goodwin (1999) found that judges were quite good at predicting borderline performance. In her study, the mean difference between estimated and empirically calculated  $p$ -values for items was only 0.03. Impara and Plake (2001) reported similar results. Others have found that judges have difficulty with this task (Fehrmann et al., 1991; Norcini, 1994; Impara & Plake 1998). Impara and Plake (1998) reported that while judges were better at estimating item-level performance for the overall group than for the minimally acceptable examinees, they were not particularly good at either task. Norcini (1994) found that judges in another experiment self-reported difficulty in predicting how borderline candidates would perform on an individual test

item. Simplifying the process so that judges estimate performance for only one candidate using a yes/no format has produced mixed results (Impara & Plake, 1997; Chinn & Hertz, 2002).

The selection procedure for judges must therefore not only consider judge expertise, but also their anticipated ability to conceptualize a pool of borderline examinees. Judge training, including the opportunity for practice, becomes critically important, since it contributes to the ability of judges to complete this task (Kane, 1994; Hambleton & Plake, 1995). Plake and Impara (2001) hypothesized that poor results of their previous study (1998) were likely attributable to a lack of judge training and practice.

Another concern related to the concept of minimal competence is the issue of conceptual drift during the standard setting process. Do the judges' concept of minimal competence remain the same over the entire standard setting process, or is it influenced by factors like exposure to test items, panel discussion, or fatigue? Drift is a potential problem in any judgmental method, but particularly when the cognitive demands of the cut score setting task are high.

*Dimensionality of the test.*

For a total test score to be a valid criterion for mastery, an assumption must be made that the construct underlying the test is either a) unidimensional, or b) multidimensional but the dimensions are compensatory in nature. Hambleton and Plake (1995) cite this as a potential weakness of applying relative weights in the Angoff procedure, though this criticism can be extended to all Angoff methods. Current Angoff methods calculate cut scores that do not account for which questions were answered

correctly or incorrectly. An examinee can meet a performance standard set using an Angoff method either by being minimally competent on all dimensions or areas of a test, or by making up for deficiencies on a given dimension with strengths in other dimensions. Using an Angoff method, judges only decide the *probability* that a minimally acceptable candidate will answer an item correctly, but they cannot determine that a student *must* answer a question correctly in order to be considered competent. Even if an item is given a *p*-value (approaching 1.0), implying that borderline candidates are very likely to answer the question correctly, the final cut score set does not specify which, if any, items must be answered correctly to reach that cut score. This deficiency of the method can be overcome by calculating separate cut scores for each dimension on a test, with examinees required to surpass each of them in order to meet the standard of competency. However, the decision to create separate dimension cut scores is contingent on whether a compensatory or conjunctive model is appropriate.

#### *Interpretation of the Cut Score*

Standards are supposed to represent the ideation of the goals of a program of study or set of criteria into cut scores on a test (van der Linden, 1982), which are in turn translated and operationalized in a cut score. In reality the true relationship of the performance standard and the cut score is not known (Woehr et al., 1991). The judgment that is involved is value-laden, and it is important for standard setters to be aware of the values they use when arriving at standards. The cut score from any judgmental method is likely to be adjusted based on the political, economic, social or educational implications of the decisions that are made based on the standard (Berk, 1995). Regardless of methods

that are used, the process needs to be documented in such a way that the logic behind decisions made is understandable and defensible to the stakeholders.

### *Precision of the Cut Score Setting Process*

When decisions are made using standards, both errors associated with the observed test score and the position of the cut score can contribute to incorrect classification of candidates. Brennan and Lockwood (1980) assume these two errors were uncorrelated. Kane and Wilson (1984) dispute this idea, since the test scores and the cut scores were based on the same set of test items. Instead, they argue that the covariance of the item main effects (from G-theory, Brennan, 2001) of measurement and the cut score is an important barometer of how well the standard setters are specifying the criteria for the standard in relation to the construct being measured by the test. A negative covariance between item errors would imply a mismatch between the criteria the judges are using and the test construct. A mismatch would greatly increase the error in the positioning of the cut score.

Kane and Wilson (1984) also suggest that the greater the agreement between item estimates and actual performance data, the lower the error associated with the cut-off score, as well as the lower the rate of examinee misclassification (Norcini et al., 1998). This is because the probability of a misclassification is an increasing function of total error variance (Kane & Wilson, 1984).

When different panels examined the same items in two separate years, the item performance estimates using Angoff methods were very similar (Impara & Plake, 2001). Further the inter-rater reliability between years was as high as the intra-judge reliability within years (Plake, Impara, & Irwin, 2000) indicating the stability of estimates using the

Angoff method over time. However, Goodwin (1999) cautions that comparing  $p$ -values of a total population to cut scores as a measure of intra-judge reliability is misleading, and suggests it is better to look only at borderline examinees, since that is what the MPLs of performance are based on. However, often performance information specific to minimally acceptable candidates is not available (Plake, Melican, & Mills, 1991). Further, correlations are not a measure of the actual goodness of fit (or precision of agreement), only an indication of the direction of the relationship between the variables (Impara & Plake, 1996; Plake et al., 2000).

It is important to consider standard error of measurement when setting cut scores. Woehr et al. (1991) compared seven different standard setting procedures and found that while they all produced different cutoff scores, all of the scores fell within the standard error of measurement of each other. However, content-based standard setting procedures (including Angoff) produced greater numbers of incorrect classifications.

The need for high inter-rater and intra-rater reliability in order to validate the Angoff method raises a conundrum: what if good measures of reliability are the result of the procedure (e.g., introducing normative data, discussion of results among panelists in an iterative process) and no longer reflect a judge's true perceptions or expectations about examinee behaviour (Plake et al., 2000). The essence and *raison d'être* of a judgmental process would be lost if the opinions of the individual judges are lost in the process.

#### The Criteria for Assessing Cut Score Methods

Berk (1986) developed fifteen criteria to assess cut score setting methods. Six of the ten common criteria were developed to assess the technical sufficiency, based on: a) standards and recommendations made by the American Educational Research

Association (AERA), American Psychological Association (APA), and National Council of Measurement in Education (NCME) Joint Committee in *Standards for educational and psychological research* (1985), b) then-current expert opinion of standard-setting researchers, and c) pertinent legal decisions. These criteria reflect the state of measurement research, as well as the political climate that surrounded standard and cut score setting at the time. The four remaining common criteria were used to assess what Berk termed the “practicability”, or how feasible a method was to put into use. The need for this type of criteria reflects the truism that a standard setting method is only as good as it is useful in a practical setting. In addition to the ten criteria used for all standard setting methods, Berk (1986) included five criteria that were specific to judgmental standard setting methods. The need for additional assessment criteria for judgmental methods is a manifest of not only the added dimension or depth that expert judges bring to cut score setting, but also to the added complexity that is part and parcel with human judgment. Subsequently, Hambleton (2001) noted twenty evaluative questions that are specific to judgmental methods. The answers to these questions reflect the quality of the cut score setting procedure, and are more comprehensive than Berk's judgmental criteria. The Angoff procedure and its modifications, as outlined above will be assessed according to both Berk's (1986) ten general criteria and Hambleton's (2001) twenty judgmental criteria. Recommendations for using the Angoff procedure, in response to the findings from the application of these criteria, will also be made.

*Berk's Criteria**Technical adequacy.*

- 1) *The method should yield classification information in dichotomous or polytomous form, as appropriate.*

Berk (1986) acknowledged that in most models (including Angoff), the underlying assumption is that the distribution of ability is continuous, but that decisions must be made to determine where the line between master/non-master or standards of excellence/proficiency/below proficiency. The cut scores must be set so that examinees can be placed clearly and unambiguously into a category based on their score. The Angoff method, with or without any of the potential modifications outlined above adequately meets this criteria because the end product of the process is an MPL or MPLs, which separate the pool into groups.

It would be prudent to include a 'region of indecision' (perhaps if both error of measurement and the stakes of the test are expected to be high), where a gap of 1 standard error of the test scale separates the performance standards. Examinees who fall into this region would be retested until their scores fall into either one of the categories decisively. However, the feasibility of this practice may be limited.

- 2) *The method should be sensitive to examinee performance.*

The cut-off score(s) should be realistic, and reflect the performance requirements of the examinees in the context under which they usually perform, not just the test setting. The use of impact data may be deleterious in this regard, since it only reflects test performance.

In addition, it should be explicitly stated for which population of examinees (and the scope and bounds of that population) the cut scores are intended. The use of judges who have expertise not only with the content area of the test, but also with the pool of examinees is important to ensure that this criteria is met. While this point is not discussed in any of the modifications listed above, it is recommended that the judges' experience be considered when selecting judges to use the Angoff procedure.

*3) To ensure fairness of the test, the method should take into account only what examinees were given the opportunity to learn.*

This criteria should be considered by test developers a priori, but standard setters must also be vigilant in ensuring that the standards examinees are expected to meet are a reflection of the learning process, even if that process was inadequate in some way. As in criteria 2) above, this criteria can be addressed by careful selection of the judging panel, with special attention to their knowledge of the examinee population.

*4) The method should be statistically sound.*

All statistics should be calculated and interpreted correctly. There is no particular criticism or concern with the Angoff method in this regard. However, simpler statistics are easier both to calculate and interpret by all stakeholders, and thus are less prone to problems in this area.

*5) The reliability of tests, and in particular the error of measurement, should be considered when establishing the cut scores.*

Error of measurement in the cut score region(s) should be reported. Kane and Wilson (1984) suggested looking at sources of variance and to use an approach that

can identify sources of error in estimating the difference between the observed test score and the cut score.

Berk (1986) further recommends that the method should identify a standard on the “true” scale as opposed to the observed test scale. This suggestion seems to be influenced by van der Linden (1982, 1984) who argued that the method yields standards on the true score scale. An underlying assumption of his argument is that the standard corresponds to the expected score on an infinite universe of test items (or all possible forms of the test). However, standards are generally set for one form of a test at a time. The Angoff method examines only the test items available and can therefore only work on the observed score scale (Livingston & Zieky, 1989).

6) *The method should yield convergent and divergent validity evidence.*

The performance standards and corresponding cut scores should be defensible in terms of their relation to other appropriate performance variables, perhaps on another test or, when possible, in relation to performance in ‘real life’ settings. The method should also yield estimates of the probability of correctly and incorrectly classifying examinees when such performance information is available. No specific recommendations for presenting this type of information for the Angoff procedure have been made in the literature. Convergent and divergent validity evidence should be collected and assessed when conducting the Angoff procedure.

#### *Practicability*

7) *The method should be simple to implement, and should not take an unreasonable amount of time to complete.*

All steps in the process should be systematic, including the presentation of data, if applicable. This point is especially critical in judgmental methods, where it is important for the judges to understand the method they are using. The Angoff procedure is relatively straight forward, with the potential exception of the cognitive demands of determining the item  $p$ -values. In this regard, the yes/no procedure proposed by Impara and Plake (1997) makes implementing the methods simpler and more feasible.

8) *The computations involved with the method should be simple to compute.*

Berk (1986) makes the specific recommendation that the standard should be computable with either a hand calculator or computer statistical program. With the exception of using IRT parameter estimates, which requires a specialized program, all calculations for the Angoff procedure could easily be computed using a calculator.

9) *The method should be defensible and 10) credible to laypeople.*

Ultimately, the stakeholders (e.g., examinees, test sponsors, politicians, tax payers) will judge the appropriateness of the standards that are set. The method should be logical and easily understandable to the lay public, so that the process can be evaluated. Methods that are statistically "magical" in their procedures are more difficult to understand therefore to have credence. One of the greatest strengths of the Angoff procedure is that it is relatively simple to understand.

#### *Hambleton's Judgmental Criteria*

In his chapter, Hambleton (2001) frames his questions in the past tense, as an evaluation of a method that has already been used. However, if there is to be any forward

movement in improving cut score setting practices, these criteria must be used to evaluate the planned cut score setting exercise, *prior* to execution.

*11) Is the method for selecting judges defensible?*

Judges should be selected on the basis of expertise in the area being examined, along with having extensive (and preferably expert) familiarity with the population being examined, in order to be able to accurately conceptualize the behaviour of minimally competent candidates, as well as predict their item performance (Impara & Plake, 1998). For a comprehensive list of characteristics that define expert judges, please refer to Jaeger (1991). It is important that judges be representative of regional socio-demographic, and any other variables salient to the population of examinees, as required. Judges must also have a good understanding of how examinees *think* during a testing situation. More importantly, it is necessary to avoid selecting convenience samples of judges.

While the selection of judges was not addressed in Angoff's original description, subsequent users of the procedure have been careful to select appropriate judging panels.

*12) Are there sufficient numbers of judges both to ensure that the panel is representative of expert opinion in the field being studied, as well as to ensure that no particular judge's scores unduly influences the cut score that is set?*

Jaeger (1991) recommended that at least 15-20 judges are necessary when setting standards for state-wide testing. Similar numbers are likely to be needed in a provincial setting.

*13) Will two panels be used to check the generalizability of the performance standards?*

While the expense of two judging panels can be prohibitively expensive, a judging panel can be divided into two randomly equivalent groups and independently determine cut scores. This step provides the basis for estimating standard error of the process (Hambleton, 2001). The Angoff procedure, in its current form, is flexible enough to allow for this step.

*14) Will sufficient resources be allocated to carry out the study properly?*

The Angoff procedure (and all judgmental methods) are costly relative to non-judgmental methods because they require large amounts of time from qualified judges. Other expenses like travel, hosting and accommodation, as required, make this type of method very expensive. Steps that add time to the process, including adequate judge training time, will add to this expense.

*15) Will the method be field tested in preparation for use in the actual cut score setting study?*

No specific recommendation for Angoff procedure applies. However, any set of procedures would ideally be field tested prior to use in a high stakes scenario.

*16) Is the cut score setting method appropriate for the particular educational assessment?*

The Angoff method was designed for and is best suited to multiple choice formats.

*17) Will panelists be explained the purpose of the assessment and use(s) of the test score at the beginning of the process?*

The judges will need to have this purpose and use in mind when determining the meaning of minimal competence, which will in turn inform the final cut score. It is

very important for users of the Angoff procedure to keep in mind the ultimate purpose of the standard.

*18) Will the qualifications and other pertinent panelist data be collected?*

This question is closely related to the discussion about point 11).

*19) Will the judges take the test before the standardizing procedure begins so that they have a better understanding of what the examinees experienced when they wrote the test?*

Particularly in light of point 11), it is critical that the judges understand the cognitive processes and strategies that examinees used during the test. One of the weaknesses of the current Angoff method is that discussing the rationale for why minimally competent candidates would be likely (or unlikely) to answer a question correctly will be fairly limited, especially with tests that have many items. Taking the test themselves is one way for the judges to better understand the thought processes of test-takers.

*20) Will the judges be trained in the method so that they have a clear understanding of their objectives and the proper process they are to follow?*

Because the Angoff method can be a cognitively taxing task, judges need the opportunity to practice and become comfortable with the procedure before beginning the item evaluation process. This practice time also allows time for judges to ask questions and gain clarification of the precise method they are supposed to follow.

*21) Will the judging panel develop clear descriptions of the behaviors associated with each category of proficiency?*

The existence of a clear performance standard makes the process of establishing cut scores easier (Kane, 2001). This point is especially pertinent to the Angoff procedure where the judges must be able to visualize a pool of candidates who minimally meet the standards.

*22) Will a moderator be used to help the judging panel discuss and reconcile differences?*

The role of a good moderator in panelist discussions cannot be underemphasized. Both Berk (1986) and Hambleton (2001) warned against the problem of social comparison, and excessive influence of judges with particularly strong points of view. Both researchers argue that it is important to have open discussion without introducing bias to the process, or preventing judges from coming to decisions independently. Close attention to the selection of judges will help in this regard.

*23) Will the process run efficiently?*

Hambleton emphasizes the need for the cut score setting process to be as straightforward and time-effective as possible, so as to minimize the burden on judges. In the Angoff procedure, which is highly cognitively demanding, this consideration is paramount. Any administrative steps that can be taken to simplify and streamline the job of the judges (without sacrificing the quality of their judgments) and reduce potential fatigue factors will be important to improving the process.

*24) and 25) Will test result or impact data be introduced to the judgmental process so that judges can observe how their cut scores behave in practice?*

The Angoff modification listed above addresses this issue adequately. The presentation of this type of data should come at the end of the cut score setting process,

so that it does not unduly sway the opinions of the judges. Judges should also be given instruction on how to use these data. In the Angoff procedure, it is important that overall impact data do not change the judges' conception of minimal competence.

*26) Will the approach for arriving at final performance standards be clearly described?*

The cut score setting process must be clearly explained to judges (as part of their training, see criteria 20) ). The process must also be documented and understandable to stakeholders not directly involved in the cut score setting exercise. This point echoes Berk's tenth criteria.

*27) Will an evaluation of the process be carried out by the judges?*

When the judges evaluate the process and find it to be satisfactory, it gives credibility to the final cut scores that are established.

*28) Will validity evidence be gathered? What form will it take?*

The performance standards must be validated. As it relates to establishing the cut score in the Angoff procedure, particular attention to the validity of the conception of minimal competence, and the rationale for item-level judgments will be important. Because the Angoff method has so many permutations, it is also important that the choice of procedures be defensible.

*29) Will the full standard setting process be documented?*

A technical report with all information and answers to the previous questions should be developed as part of the validity evidence.

*30) Will effective steps be taken to communicate the performance standards?*

Hambleton suggests the use of exemplars as a possible way to communicate the meaning of membership in a particular group (e.g., a clear description of the

performance that defines "competent") to the stakeholders and the public. In the Angoff procedure, this is a step that might be helpful not only to help stakeholders understand the process afterwards, but also to the judges themselves, prior to the standard and cut score setting exercise.

The overarching principle guiding these criteria is developing the validity of the cut score and ensuring that the interpretation of the cut score (and by extension, the standard) is valid for the context in which it will be used. Criteria 5) and 18) refer specifically to the gathering of validity evidence, but all of these criteria contribute to the overall validity of the standard setting process. It is important to consider that validity is not a binary outcome; that is, validity varies in degree. Trying to maximize the validity of a procedure is a never-ending process. The procedures used for establishing the defensibility of the method will depend not only on the context, but the method itself. Certainly the criteria above will have trade-offs with each other in some cases. It is likely that the technical adequacy of a method will have to be balanced with practical considerations. The importance of each criteria will vary from testing scenario to testing scenario, and in some cases not all of the criteria listed above will be appropriate. The users in each scenario will ultimately have to decide what will work best in their own case. As Jaeger (1991) and Kane (1994) pointed out, while cut scores and standards are judgments and therefore arbitrary, they can still be legitimate if the procedure used to determine the cut score is logical and steps to minimize subjectivity in the process have been taken.

*Overall Performance of the Angoff Procedure*

The main strength of the Angoff procedure is its simplicity. This method is relatively simple to explain to judges and to stakeholders. It uses simple statistics that are easy to compute and understand. It is not a "magical" procedure. Its main weakness is its deceptively high cognitive load for judges, particularly if there are many items on a test. Modifications to simplify the task for judges, without risking the quality of the judgment, will help improve this weakness.

### Conclusions

The attention and resources currently being devoted to testing leads to increasing importance of the assessment of the validity and defensibility of standard-setting methods. The criteria compiled here reflect the need to validate any standard setting procedure, and further emphasize the importance of the selection and preparation of the judges to conduct the standard setting exercise.

The modifications to the original Angoff procedure discussed here have mixed results. In general, the modifications are not always suitable for every situation, but they are useful and will improve the procedure when they are appropriate. In particular, the *judicious* use of normative and impact data and group discussion during an iterative process are important contributions to increasing the validity of the Angoff procedure.

The Angoff method continues to be plagued by some nagging criticisms. The ability of the judges to conceptualize a minimally competent candidate is a problem that is difficult to overcome. Emphasis on clearly defining what a borderline candidate is, as well as judge training will improve the outcomes of any Angoff procedure. Additionally, whenever possible, outcomes of standard setting should be validated by comparing it to empirical data.

There is a certain irony in developing standards to evaluate cut score setting techniques. It can almost be considered an infinitely circular process when one sets out to evaluate the proficiency of techniques that are used to define proficiency. Like any other set of criteria, these criteria for evaluating cut score setting methods will need to be continually updated, to reflect the progress made in research, and the social and political climate in which test standards are set.

## References

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) Joint Committee. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.
- 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* (2<sup>nd</sup> ed., pp. 508-600). Washington, DC: American Council on Education.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137-172.
- Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do! *Applied Measurement in Education*, 8, 99-109.
- Berk, R. A. (1996). Standard setting: the next generations (where few psychometricians have gone before!). *Applied Measurement in Education*, 9, 215-235.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, 4, 219-240.

- 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.
- Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education, 12*, 151-165.
- Chinn, R. N., & Hertz, N. R. (2002). Alternative approaches to standard setting for licensing and certification examinations. *Applied Measurement in Education, 15*, 1-14.
- Fehrmann, M. L., Woehr, D. J., & Arthur, W. (1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. *Educational and Psychological Measurement, 51*, 857-872.
- Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education, 12*, 13-28.
- 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.
- Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41-55.
- Impara, J.C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353-366.

- Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69-81.
- Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3-6, 10.
- Kane, M. T. (1987). On the use of IRT methods with judgmental standard setting procedures. *Journal of Educational Measurement, 24*, 333-345.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.
- Kane, M. T. (2001). So much remains the same: Conception and validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum Associates.
- Kane, M. T., Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement, 8*, 107-115.
- Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*, 121-141.
- Norcini, J. J. (1994). Research on standards for professional licensure and certification examinations. *Evaluation and the Health Professions, 17*, 160-176.
- Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25*, 57-65.
- Plake, B. S., & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment, 7*, 87-97.

- Plake, B. S., Impara, J.C., & Irwin, P.M. (2000). Consistency of Angoff-based predictions of item performance: Evidence of technical quality of results from the Angoff standard setting method. *Journal of Educational Measurement, 37*, 437-355.
- Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational measurement: Issues and Practice, 10*(2), 15-16, 22, 25.
- Reid, J. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues & Practice, 10*(2), 11-14.
- Taube, K. T. (1997). The incorporation of empirical item difficulty data in the Angoff standard-setting procedure. *Evaluation and the Health Professions, 20*, 479-498.
- United States Government. (2002). The facts about...Measuring Progress. Retrieved November 11, 2002, from <http://www.nochildleftbehind.gov/start/facts/testing.html>
- van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19*, 295-308.
- van der Linden, W. J. (1984). Some thoughts on the use of decision theory to set cutoff scores: Comment on de Gruijter and Hambleton. *Applied Psychological Measurement, 8*, 9-17.
- Woehr, D. J., Arthur, W., & Fehrmann, & M. L. (1991). An empirical comparison of cutoff score methods for content-related and criterion-related validity settings. *Educational and Psychological Measurement, 51*, 1029-1039.