The Alberta Journal of Educational Research *Vol. 52, No. 1, Spring 2006, 53-64*

Kathryn L. Ricker
University of Alberta

Setting Cut-Scores: A Critical Review of the Angoff and Modified Angoff Methods

The purpose of this article is to review critically the Angoff (1971) and modified Angoff methods for setting cut-scores. The criteria used in this review were originally proposed by Berk (1986). The assumptions of the Angoff method and other current issues surrounding this method are also discussed. Recommendations are made for using the Angoff method. In addition, several issues that are relevant to cut-score setting that are not addressed by Berk's criteria arose while reviewing the Angoff method. These issues are addressed separately.

Cet article a comme objectif d'examiner, de façon éclairée, la méthode Angoff (1971) et la méthode Angoff modifiée employées pour établir les notes de passage. Cet examen reprend des critères proposés par Berk (1986). On présente les hypothèses sur lesquelles repose la méthode Angoff, discute des questions relatives à la méthode et propose des recommandations pour son utilisation. Finalement, on présente plusieurs questions qui se sont avérées pertinentes à l'établissement des notes de passage mais dont ne tiennent pas compte les critères de Berk.

Introduction

There has been a decided movement toward standardized tests in North America. All provinces in Canada with the exception of Saskatchewan and Prince Edward Island use some sort of standardized provincial examination. All provinces and territories participate in the national Pan-Canadian Education Indicators Program (PCEIP, formerly SAIP) test. Canadian students also participate in international tests including the Programme for International Student Assessment (PISA).

Without question, the largest-scale impetus toward standardized testing can be witnessed in the United States with Public Law 107-110, *No Child Left Behind Act* of 2001 (United States Government, 2002). This Act includes assessment and evaluation of student progress as one of four main pillars. As a result, each state is federally mandated to develop statewide tests of student achievement in core curriculum areas (mathematics and reading currently, science by 2005-2006) in grades 3 to 8. Over $410 million in funding has been allocated to aid states in the development and administration of these tests for 2005 alone.

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 of the learning that the tests are designed to assess. Ideally this performance standard is the embodiment of the learning objectives for the students being assessed. The standard should represent mastery of the

Kathryn Ricker is an associate measurement statistician in the Center for Statistical Analysis at the Educational Testing Service in Princeton, New Jersey. At the time of writing, she was a doctoral student in the Centre for Research in Applied Measurement in Evaluation.

learning objectives, the level of basic proficiency necessary to move on to the next level, or the skills to function in the real world (van der Linden, 1982). In effect, establishing a performance standard can be conceptualized as policy-making that has an effect on everyone involved in the testing procedure (Kane, 2001).

Once established, a performance standard is translated into a cut-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-score is to separate examinees that meet the standard from those who do not. The distinction between establishing a performance standard and setting a cut-score is often confusing because generally, one task is not done without the other. However, it is important to keep in mind that standard-setting is a policymaking activity, whereas setting a cut-score is the operationalization of that policy. Although the Angoff (1971) and the other methods for setting cut-scores (e.g., Bookmark, Body of Work, Analytic Judgment) are often referred to as methods for standard-setting, they are actually used to set cut-scores 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. 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 relation to established standards?

Angoff (1971) initially proposed a method that involved indicating whether a minimally competent person would answer an item correctly. In a footnote he presented a variation that, using how often it is used and the amount of attention devoted to it in the research context as indicators, is one of the most commonly used methods for setting cut-scores today. The variation, which many people take to be the original method, has been modified in various ways (Hambleton & Plake, 1995; Impara & Plake, 1997; Taube, 1997) in attempts to improve it.

In this article the variation presented by Angoff (1971) is referred to as the *basic Angoff method*. The basic method and four modifications made to how it has been used are evaluated in terms of the 10 criteria established by Berk (1986). Also included in the article are a synthesis of some of the empirical investigations of Angoff-type methods conducted since 1986 and a discussion of the current debates and issues relevant to the Angoff methods. Some recommendations for additional modifications of the Angoff method that best meet Berk's criteria are made.

The Angoff Methods

Basic Angoff Method

The basic Angoff method appears to be a simple process. A group of judges are asked to think of a group of minimally competent candidates who would border on the mastery or non-mastery cut-off score. The most typical instruction is for judges to think of a pool of 100 candidates who would just barely meet the performance criteria. The judges, working independently, then estimate what percentage of the sample of minimally competent candidates would correctly answer each item in the test. The percentages are converted to

a *p*-value for each item. These *p*-values are summed to form the Minimum Passing Level (or cut-score) for each judge (MPL_j, j=1…J, where J is the number of judges). The mean of these cut-scores is the final cut-score or the *MPL* for the test. The standard error of the final cut-score is equal to the standard deviation of the J judges' *MPL*s. A low standard error is desirable because it denotes better agreement among the judges and less uncertainty about where the true cut-score should lie.

This method does not apply only to setting cut-scores to separate minimally competent from non-minimally competent candidates. It can also be used to create a cut-score for any grouping in the population or multiple cut-scores. For example, the basic Angoff method 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 and set a cut-score to separate them from the remaining examinees.

Modifications of the Basic Angoff Method

In an attempt to improve the agreement among judges, several modifications and interpretations have been made to the basic Angoff method. These modified procedures are frequently referred to using the general term *modified Angoff* procedures. Presented below is a discussion of the effects of four specific modifications that have been empirically tested and reported in the scholarly literature.

Using an iterative process. Conducting a number of rounds or iterations of the cut-score-setting process is the most commonly used modification of the basic Angoff procedure (Busch & Jaeger, 1990; Hambleton & Plake, 1995; Jaeger, 1978; Woehr, Arthur, & Fehrmann, 1991). Iteratively setting cut-scores is listed as a desirable characteristic of a judgmental process (Hambleton, 2001). The time between iterations or 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 (Chang, 1999) or three (Busch & Jaeger) iterations has been reported in the literature. Busch and Jaeger found that an iterative process reduced variability among the judges item estimates and consequently their cut-scores.

Presentation of normative data. The second type of modification involves providing performance data to the judges. Like the first, the intent of this modification is to improve interjudge agreement. Data are presented to the judges before the final iteration. Busch and Jaeger (1990) found that presentation of item difficulties to the judges resulted in an increase in the correlation between item *p*-values and item difficulties, as well as greater agreement among the judges (i.e., increased interjudge reliability). Norcini, Shea, and Kanya (1988) found that judges used normative data about 25% of the time such data were presented and that the average change in the estimated *p*-value per item was relatively small, with a tendency for the change to occur on items that had originally been estimated by the judges 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. As indicated above, Angoff first suggested that judges determine the probability of a single minimally competent candidate correctly answering an item. But as indicated above, estimating the proportion

of minimally competent examinees who would correctly answer a test item, which was presented in the footnote to the first suggestion, is the procedure commonly used. However, Nassif (1978), seemingly unaware of Angoff's first suggestion, proposed that judges decide whether a single minimally competent candidate would or would not answer an item correctly. Jaeger (1978) proposed a similar idea, combining the yes/no judgment with the use of an iterative process. The rationales for this modification are that it is easier for judges to think of a single person than a pool of candidates and to make a simple yes/no decision.

The results of empirical investigations of this modification are equivocal. Impara and Plake (1997) compared the traditional percentage method with the yes/no method. They reported that the two methods produced essentially the same cut-score and that the iterations produced only fine tuning of the initial estimates. They postulated that although individual judges may not be accurate at predicting individual performance, judges as a group produce a reasonable cut-score when aggregated. However, Impara and Plake also conceded that their results might have been contaminated by the panel using the percentage method immediately before using the yes/no method.

Chinn and Hertz (2002) compared these same methods in experiments using two separate test forms. Two groups of judges standard-setting used the percentage method; two groups of judges used the yes/no method. Judges reported that the yes/no procedure was easier to conduct, but their final cut-scores were far more influenced by the empirical data with which they were presented, and their overall cut-scores were less stable over iterations of the process when compared with the percentage method.

Applying relative weights to scores. Hambleton and Plake (1995) developed what they referred to as the "extended Angoff" procedure. They applied Angoff-type procedures using a four-point scale to multidimensional, polytomously scored exercises. In addition to predicting scores of minimally competent candidates on each dimension, raters provided a weight for each dimension and each exercise, where the weights reflected the relative importance of the content of each dimension to the exercise and of each exercise to the total test content. When evaluating this method, judges were more confident in their ratings. Furthermore, they had greater confidence in the ratings than when the dimensions were not weighted.

The Criteria for Assessing Cut-Score Methods

Berk (1986) developed 10 criteria to assess cut-score setting methods. Six were developed to assess the technical sufficiency of the method. The technical criteria were based on: (a) standards and recommendations made by the American Psychological Association (APA), American Educational Research Association (AERA), 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 reflected the state of measurement research and the political climate that surrounded establishing standards and setting cut-scores at the time. The four remaining criteria were used to assess what Berk termed the *practicability*, or how feasible a method was to put into use.

Using a three-point scale, the basic Angoff method and its variants where appropriate are rated for how well they meet Berk's criteria (1=not met, 2=partially met, 3=fully met). Recommendations for using the Angoff procedures in response to the findings from the application of these criteria are also made.

Technical Adequacy

1. The method should yield appropriate classification information (Rating: 2). The Angoff method, with or without any of the modifications outlined above, adequately meets this criterion because the end product of the process is an MPL, which separates the examinees into the varied but adjacent performance levels. More precisely, the method should yield classification information in dichotomous or polytomous form (depending on the number of cut-scores set) as appropriate. The underlying assumption of the Angoff procedures is that the distribution of ability is continual, but that decisions must be made to determine where the cut-score between master/non-master or the cut-scores between excellence, proficiency, and below proficiency lie. The cut-scores must be set so that examinees can be placed clearly and unambiguously into a category based on their score.

When various panels examined the same items in two separate years, the item performance estimates using Angoff methods were similar (Plake & Impara, 2001). Furthermore, the interrater reliability between years was as high as the intrajudge reliability within years (Plake, Impara, & Irwin, 2000) indicating the stability of estimates using the Angoff method over time. Goodwin (1999) cautions that comparing *p*-values of a total population to cut-scores as a measure of intrajudge reliability is misleading and suggests that it is better to look only at minimally competent examinees because this is what the MPLs of performance are based on. However, often performance information specific to minimally competent candidates is not available (Plake, Melican, & Mills, 1991). Furthermore, 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 that are correlated (Impara & Plake, 1997; Plake et al., 2000).

The need for high interrater and intrarater reliability in order to validate the Angoff method presents 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 behavior. The essence and *raison d'être* of a judgmental process would be lost if the opinions of the individual judges were lost in the process.

2. The method should be sensitive to examinee performance (Rating: 2[3]). The basic Angoff cut-scores (without the presentation of impact data) are intended to correspond to a standard of performance that is independent of student performance on the test. However, cut-scores must also be realistic. Without the presentation of impact data, there is no way of assessing if cut-scores are realistic. Presenting impact data such that it balances the need for a sense of realism with the need for standards (and by extension, cut-scores) to retain their intended meaning when conducting the Angoff method with impact data would change the rating for this criterion from 2 to 3.

3. The method should be sensitive to instruction or training (Rating: 1). No specific guidelines or references to this issue are made in the Angoff methods. As with the second criterion, this criterion can be addressed by careful selection of the members of the judging panel, with special attention paid to their knowledge of the learning experiences of the examinee population. To ensure fairness of the test, the method should take into account only what examinees were given the opportunity to learn. This criterion should be considered by test developers a priori, but judges must also be vigilant in ensuring that the performance standards examinees are expected to meet are a reflection of the learning process.

4. The method should be statistically sound (Rating: 3). There is no particular criticism or concern with the Angoff method in this regard. Simple statistics are easier to interpret by all stakeholders, but it is important to ensure that the statistics adequately capture the notion of the cut-score.

When decisions are made about cut-scores, both the 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 are uncorrelated. Kane and Wilson (1984) dispute this idea because the test scores and the cut-scores are based on the same set of test items. Instead they argue that the covariance of the item main effects and the cut-score (from G-theory, Brennan, 2001) is an important barometer of how well the judges estimate the cut-score. Whereas a positive covariance suggests good criteria-construct fit, 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 is the error associated with the cut-score, and consequently, the lower is the rate of examinee misclassification (Norcini et al., 1988). This is because the probability of a misclassification is an increasing function of total error variance of the cut-score.

There is some controversy over whether the use of standard error is appropriate to describe the variability in cut-scores in the pool of judges. MacCann and Stanley (2004) argue that the standard error is inappropriate because the judges' ratings are not independent after the first iteration. This dependency violates a postulate of the Central Limit Theorem, which requires random, independent sampling. The Central Limit Theorem acts as the link between the observed standard deviation and the calculation of the standard error. However, it can also be argued that how the judges are introduced to the Angoff method to be used and the common training they receive would also create dependency before the judging process, or that any dependency created by judge discussion between iterations can be diminished by asking judges to make their decisions independently of each other, using their own judgment with the extra information of the iterations providing guidance. Currently the *Standards for Educational and Psychological Testing* call for reporting "an estimate … of the amount of variation in cut scores that might be expected if the standard-setting procedure were replicated" (AERA/APA/NCME, 1999, Standard 4.19, p. 60). More research in this area is required to clarify this issue.

It is important to consider the standard error of measurement when setting cut-scores. Woehr et al. (1991) compared Angoff and six other cut-score-setting procedures and found that although they all produced varying cutoff scores, all scores fell within the standard error of measurement of each other. However, content-based (versus criterion and normative) standard-setting procedures (including Angoff) produced greater numbers of incorrect pass/fail decisions.

5. The method should identify the true standard (Rating: 2). The Angoff method does not directly address this issue in the original method or any of the variations to the method listed above, but could easily be incorporated. Berk (1986) rated the Angoff method as having a "marked advantage" (p. 155, Table IV) with regard to this criterion, perhaps because the method could be easily adapted to include consideration of error of measurement in the region of the cut-score.

The dependability of tests and in particular the absolute error of measurement should be considered when establishing the cut-scores. The absolute error of measurement in the cut-score region(s) should be reported. Kane and Wilson (1984) suggested looking at sources of error variance and using 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 cut-score 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 should yield cut-scores 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, cut-scores 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).

It would be prudent to include a *region of indecision* where a gap of one standard error of measurement in the test scale separates the performance standards. This practice takes into account that measurement is prone to error and that most tests have less than perfect dependability. Examinees who fell into this region would be retested until their scores fell into either one of the categories decisively. However, the feasibility of this practice is probably limited, and it is not necessary unless the standard error of the cut-score is large.

6. The method should yield decision validity evidence (Rating: 2). The Angoff method does not directly address issues of validation. Berk's criterion specifically mentions that the method should include estimates of the probability of classification errors, which is not part of the Angoff methods considered in this article. However, convergent and divergent validity evidence can and should be collected and assessed when conducting the Angoff procedure. The performance standards and corresponding cut-scores should be defensible in terms of their relationship to other appropriate performance variables, perhaps on another test of the same construct (achievement), 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. Comparing *p*-values obtained empirically to the judges' ratings provides one form of validity evidence using the IRT method outlined by Taube (1997). However, this type of evidence would be inadequate validation by itself.

Practicability

7. The method should be easy to implement (Rating: 2). The Angoff procedure seems relatively straightforward, with the potential exception of the cognitive demands of the method on the judges. In this regard, the yes/no procedure proposed by Angoff (1971) and later by Nassif (1978) (see also Chinn & Hertz, 2002, Jaeger, 1978) makes implementing the methods simpler. Steps in the process should be systematic, including the presentation of data. This point is especially critical in judgmental methods where it is important for the judges to understand the method they are using. The method should also be such that it can be completed in a reasonable amount of time so that judges are not unduly burdened.

One of the main assumptions of the Angoff method(s) is that a pool of expert judges will be able to conceive of a comprehensive and appropriate picture of what minimally competent 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 minimally competent examinees as a "nearly impossible cognitive task" (p. 216), whereas Shepard (1995, cited in Plake & Impara, 2001) claimed that the Angoff 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 minimally competent performance. In her study the mean difference between estimated and empirically calculated *p*-values for items was only 0.03. Plake and Impara (2001) reported similar results. Others have found that judges have difficulty with this task (Fehrmann, Woehr, & Arthur, 1991; Norcini, 1994; Impara & Plake 1998). Plake and Impara reported that although judges were better at estimating item-level performance for the overall group than for the minimally competent examinees, they were not particularly good at either task. Norcini found that judges self-reported difficulty in predicting how minimally competent 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, consider not only judges' expertise, but also their anticipated ability to conceptualize a pool of minimally competent examinees. Training of judges, including the opportunity for practice, becomes critically important because 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 earlier study (Impara & Plake, 1998) were probably attributable to a lack of judges' training and practice.

Another concern related to the concept of minimal competence is the issue of conceptual drift during the standard-setting process. Do judges' concepts of minimal competence remain the same over the entire standard-setting process,

or is their conception 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.

The yes/no procedure warrants further research. The idea is to simplify the decision-making process that judges must engage in for each item. However, some evidence suggests 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 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, stable, and defensible result. If the yes/no method is not comparable to the traditional *p*-value estimation procedure, then perhaps it would be better to increase judges' training and practice to help judges manage the process, rather than trying to make the method simpler.

8. The method should be easy to compute (Rating: 3). Berk (1986) makes the specific recommendation that the cut-scores should be computable with either a hand-held calculator or computer statistical program. All calculations for the Angoff procedure can easily be computed using a calculator.

9. The method should be easy to interpret (Rating: 3). The Angoff methods produce resulting cut-scores that are easily understandable.

10. The method should be credible to laypeople (Rating: 3). One of the greatest strengths of the Angoff procedures are that they are relatively simple to understand. Methods that are statistically "magical" in their procedures are more difficult to understand and therefore to have credence.

Standards are supposed to represent the goals of a program of study (van der Linden, 1982), which in turn are 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 made in determining cut-scores is value-laden, and it is important for standard-setters to be aware of the values they use during the process. Further, the cut-score from an Angoff 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). Ultimately, the stakeholders (e.g., examinees, test sponsors, politicians, taxpayers) will judge the appropriateness of the cut-scores that are set. The method should be logical and easily understandable to the lay public so that the process can be evaluated for its defensibility.

Summary

Overall, the Angoff method(s) fare well when assessed using Berk's (1986) criteria. These methods met or partially met five of the six technical adequacy criteria and all four of the practicability criteria. At the minimum the simplicity of the method makes it such that slight modifications or accommodations make it relatively simple to adapt to each specific criterion. In other words, although the described methods did not specifically meet each criterion, it is possible to adapt the method in specific cut-score-setting situations so that it does so in almost every case.

The main strength of the Angoff method (and all its modifications) is 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. The main drawback to the method is that it is cognitively burdensome to the judges, particularly if the cut-score-setting process takes a long time due to a large number of items, disagreement among the judges, or multiple iterations. The time involvement is compounded when multiple cut-scores must be determined.

Conclusions

The attention and resources currently being devoted to testing increasingly emphasize the importance of 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 produced 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 and hold the concept of a minimally competent candidate over a prolonged period is a problem that is difficult to overcome. Emphasis on clearly defining minimal competence, as well as training judges will improve the outcomes of any Angoff procedure. In addition, whenever possible, outcomes of standard-setting should be validated by comparing them with empirical data.

Overall, the Angoff method is a good method for cut-score setting, particularly when both time and resources are available to train a panel of expert judges properly. The method would be further strengthened by using Berk's (1986) criteria to guide methodological decisions that have not been explicitly addressed in the modifications to the Angoff procedure.

References

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

American Psychological Association (APA), American Educational Research Association (AERA), and National Council on Measurement in Education (NCME) Joint Committee. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological 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. (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 minimally competent 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: Erlbaum.

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. (1978). *A proposal for setting a standard on the North Carolina high school competency test*. Paper presented at the annual meeting of the North Carolina Associate for Research in Education, Chapel Hill.

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: Erlbaum.

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.

MacCann, R.G., & Stanley, G. (2004). Estimating the standard error of the judging in a modified-Angoff standard-setting procedure. *Practical assessment, research and evaluation.* Retrieved November 3, 2004, from: http://PAREonline.net/getvn.asp?v=9&n=5

Nassif, P.M. (1978). *Standard-setting for criterion-referenced teacher licensing tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto.

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*, 437-455.

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.

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). *Department of Education ESEA Title I grants to local education agencies*. Retrieved November 4, 2004, from: http://www.ed.gov/about/overview/budget/statetables/05stbyprogram.pdf

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.