2011

Using Stratified Item Selection to Reduce the Number of Items Rated in Standard Setting

Tiffany Nicole Smith
University of South Florida, tnb@mail.usf.edu

Using Stratified Item Selection to Reduce the Number of Items Rated in Standard Setting

by

Tiffany N. Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Walter Borman, Ph.D.
Robert Dedrick, Ph.D.
Stephen Stark, Ph.D.
Michael Brannick, Ph.D.
Douglas Rohrer, Ph.D.

Date of Approval:
September 2, 2011

Keywords: Performance Standard, Cut Score, Modified Angoff, Test Development, Rater Fatigue

Acknowledgements

It is a pleasure to thank those who have made this dissertation possible by generously providing me with knowledge, inspiration, and moral support. To my father, William Smith, for inspiring me by example to fulfill my every ambition, for teaching me to make every experience a learning opportunity, and for instilling in me the drive and fortitude needed to complete this process.

To my advisor, Walter Borman, for patiently allowing me to explore my research interests, for serving as a sounding board for my thoughts and ideas, and for demonstrating that scholars can use the privileges of academia to create a better world for others. To the members of my dissertation committee, Robert Dedrick, Michael Brannick, Stephen Stark, and Douglas Rohrer, for their encouraging words, thoughtful criticism, and generous time and attention to my research.

To my colleagues for sharing their enthusiasm for and comments on my work: Adrienne Cadle, Reed Castle, David Cox, Maria Gonzalez, Isaac Lee, Michael Jones, Fae Mellichamp, Christine Niero, Lynn Webb, and Cynthia Woodley.

And finally, to Wyatt Castellvi, for endless love, support, and understanding during this process. Thank you for helping me to keep my perspective on what is important in life and for providing me with a wonderful sense of anticipation for what lies ahead.

Table of Contents

List of Tables

List of Figures

Abstract

The primary purpose of this study was to evaluate the effectiveness of stratified item sampling in order to reduce the number of items needed in Modified Angoff standard setting studies. Representative subsets of items were extracted from a total of 30 full-length tests based upon content weights, item difficulty, and item discrimination. Cut scores obtained from various size subsets of each test were compared to the full-length test cut score as a measure of generalizability. Applied sampling results indicated that 50% of the full-length test is sufficient to obtain cut scores within one standard error of estimate (*SEE*) of the full-length test standard, and 70% of the full-length test is sufficient to obtain standards within one percentage point of the full-length test standard. A theoretical sampling procedure indicated that 35% of the full-length test is required to reliably obtain a standard within one *SEE* of the full-length standard, and 65% of the full-length test is required to fall within one percentage point. The effects of test length, panelist group size, and interrater reliability on the feasibility of stratified item sampling were also examined. However, these standard setting characteristics did not serve as significant predictors of subset generalizability in this study.

Introduction

Among the measurement issues examined in the criterion-referenced testing literature over the past 30 years, standard setting has received the most attention (Berk, 1986). The methods and procedures used to determine cut scores are a critical source of validity evidence for assessments designed for the purpose of selection, classification, or licensure and certification (Downing, 2006), and high-stakes decisions must routinely be made by comparing an observed or scaled test score to a pre-determined passing score. For example, a job candidate who passes a job knowledge test is assumed to have the knowledge, skills, and judgment needed to perform the job effectively. In many testing applications, candidates may not even see their actual scores, and instead may only receive information about their performance category or classification. A performance standard defines the knowledge and skill required of the examinee to demonstrate adequate competency "for the intended purposes or goals of the decision process" (Kane, 1994). Thus, every stage of the test development process relies on the accuracy and reliability of this standard, and the standard setting process has been transformed into a major component of high stakes test development (Kane, 2001).

Reviews of standard setting methods (Jaeger, 1989; Wang, Pan, & Austin, 2003; Zieky, 2001) have identified more than 50 methods for setting standards and estimating error rates once a standard has been determined. Comparisons of various methods have conclusively shown that different methods produce different standards (Berk, 1996;

Jaeger, 1989; Mehrens, 1995). In a study comparing the Angoff and Nedelsky standard setting methods, Behuniak, Archambault, and Gable (1982) similarly found that passing standards differed not only between methods, but even amongst groups of raters using the same methodology. However, none of these methods can be considered more 'correct' than other methods; the task of the subject matter experts is not to discover some 'true' passing score, but to use their professional judgment to determine the minimum performance level required of candidates to pass the examination (Downing, 2006). However, this is not to say that standards should be unsystematically set without consideration of a particular criterion. Current *Standards* (AERA, APA, NCME, 1999) concerning passing scores dictate that an absolute or criterion-referenced process should be used to evaluate the examination items, rather than setting an arbitrary cut score. Reckase (2005) argued that an appropriate standard setting method should "recover the intended standard for a panelist who thoroughly understands the functioning of test items and the standard setting process, and who makes judgments without error" (p.1). Thus, the strength of validity evidence (and legal defensibility) for a particular standard-setting method must rely on the demonstration of a reasonably unbiased process, supporting evidence for its rationale and research basis, and the psychometric characteristics of expert judgments.

Standard Setting Methodology

Standard setting methodologies are typically dichotomized into relative/normative methods and absolute methods. Relative or norm-referenced methods of standard setting are designed to pass or fail a predetermined percentage of candidates. Scores are interpreted as being better or worse than others in the norming sample, rather than as a

reflection of the candidate's level of competence (Hambleton & Pitoniak, 2006). One distinct disadvantage to normative standard setting is that the relative passing scores do not reflect the absolute competency of candidates or make any judgments about the level mastery obtained by the candidate (Downing, 2006).

In contrast, absolute passing score methods systematically use subject matter expert judgments concerning the amount of knowledge, skill, or ability required on a test to pass the examination. These methods require knowledge of both the test items and the target candidate, and can be broken into three main classifications: empirical methods, IRT-based rational methods, and classical rational methods. Empirical methods, such as the Borderline Groups method or the Comparison Groups method, use a candidate distribution on an external criterion variable to determine an appropriate performance standard (Livingston & Zieky, 1982). IRT-based rational methods, such as the Bookmark method (Lewis et al., 1996) and the Mapmark method (Schultz & Mitzel, 2005), use IRT *b* parameters to order items in terms of difficulty from easiest to hardest. Subject matter experts use the rank-ordered set of items to determine the item mapping location which separates one competency level from another (Horn, Ramos, Blumer, & Maduas, 2000).

The empirical and IRT-based rational methods require knowledge of candidates' performance data in order to compute the psychometric properties of the items. In particular, the increasingly popular IRT-based standard setting methods may require large amounts of prior performance data in order to calibrate the data, depending on the IRT model used (3PL vs. Rasch model). However, passing score studies are often conducted prior to obtaining information about candidate (and item) performance. Therefore, the focus of this research is on one of the classical rational methods, which does not require

test developers to obtain item performance information prior to setting a performance standard. Classical rational methods rely on the judgment of subject matter experts to rate each item on an examination based on estimated difficulty for minimally competent or borderline candidates in order to determine an appropriate minimum performance standard.

Nedelsky. The Nedelsky method (Nedelsky, 1954) is a classical rational method originally used primarily in education contexts to replace norm-referenced methods of standard setting (Cizek, 2006). One of the first "absolute" methods of standard setting, this method is fairly intuitive and easy to apply to multiple-choice test items. Subject matter experts are asked to individually review each item and estimate the number of response options a candidate with borderline competency would be able to rule out as incorrect. Prior to this task, the facilitator discusses and clarifies with the panelists the definition of a borderline candidate. The reciprocal of the number of choices *not* ruled out by the expert judges is used to compute the probability that the borderline candidate will answer the question correctly by guessing after eliminating all known incorrect responses options. These probability estimates are then summed across all items on the examination to compute each subject matter expert's performance standard, and these values are averaged across experts to obtain an overall cut score. This process can be repeated over multiple rounds. Although rarely used in practice, Nedelsky also suggested that the resulting cut score be adjusted for measurement error (Hambleton & Pitoniak, 2006). The recommended adjustment uses an estimate of the standard deviation of the performance standard based on the values of the reciprocals, and a constant computed using considerations of the types of items on the test.

This method has been criticized for the assumption that candidates eliminate item choices they know are wrong and choose from the remaining choices at random. Moreover, it limits subject matter experts to a discrete set of probabilities, depending on the number of response options (Brennan & Lockwood, 1980). Specifically, raters are limited to only .20, .25, .33, .50, and 1.00 for a five option item. Because a panelist would be unlikely to assign probabilities of 1.00, this could result in a lower resulting passing score than if an alternative method were used (Shephard, 1980). Some studies have provided evidence of this tendency (Chang, 1999; Melican, Mills, & Plake, 2002). The method is also limited to use with multiple-choice style examinations, which, in combination with the weaknesses previously noted, may be one reason for its decline in popularity.

Yes/No Angoff. The Yes/No Angoff method (Angoff, 1971; Impara & Plake, 1997) is the original Angoff (1971) method of standard setting, and was designed to simplify the cognitive load on subject matter experts. Following a detailed discussion of the target candidate and clarification of a minimally competent (borderline) candidate, panelists are asked to individually review each item and make judgments about whether a minimally competent candidate would answer the question correctly. Subject matter experts are only required to indicate yes or no for each item on the examination. A value of zero was provided as a rating if the panelist indicated No, while a 1 was provided if the panelist indicated Yes. Scores are averaged across items and across panelists to obtain an overall passing score for the examination.

Another variation of the Yes/No method requires panelists to make judgments with respect to an actual candidate on the borderline between passing and failing the

examination (between classifications). Conceptually, this variation can be perceived as problematic, as panelists know nothing of the passing score when referencing their judgments. Often, subject matter experts are asked to repeat and/or reconsider their original ratings after receiving feedback about actual item performance or following a group discussion of each panelist's ratings. A group consensus discussion is also frequently conducted, after which the group as a whole can alter their original averaged passing score.

As with the Nedelsky method, the utility of the Yes/No Angoff method is limited to dichotomously scored item formats. However, the potential use of this method extends beyond the Nedelsky method to include other types of dichotomously scored items (rather than just multiple-choice). In a study conducted by Downing, Lieska, and Raible (2003) comparing the Yes/No method with the Modified-Angoff and Hofstee methods, this method was found to perform reasonably well in an educational context. However, there is high potential for either negative or positive bias, because the implicit judgment of the subject matter experts is based on whether the probability of a correct response at the passing score is greater than 0.5. For example, if all of the test items were perceived by the panelists to have difficulty (p) values greater than 0.5, all ratings from an accurate subject matter expert would equal 1 (Yes), and the resulting performance standard would equal 100 percent. Certainly that would not be the intention of the panelists.

Modified Angoff. The Angoff method (Angoff, 1971) and its various modifications is one of the most popular and widely used methods of standard setting for multiple-choice examinations (Ferdous & Plake, 2007; Hurtz & Auerbach, 2003; Impara, 1995; Plake, 1998). Although the Angoff method of standard setting has been continually

adjusted throughout its history, the most commonly used Angoff variation is the Modified Angoff. William Angoff (1971) suggested the Modified Angoff method as an alternative to the Yes/No Angoff method, and this variation quickly became more widely used in practice. This method requires subject matter experts to individually review each item on an examination form, and to provide estimates of the proportion of minimally competent candidates who would correctly respond to the item. Prior to this rating task, the definition of the target candidate and the minimally competent (minimally acceptable) candidate is clarified for the subject matter experts. Proportion estimates are then summed across all items on the examination to compute each subject matter expert's performance standard, and these values are averaged across experts to obtain an overall passing score. Feedback may be provided in the form of a group discussion of individual panelists' ratings, or by sharing information about actual item performance. Following feedback from the facilitator, candidates may have the option to readjust their Angoff ratings. Many standard setting studies using the Modified Angoff technique end with a group consensus discussion, in which the group as a whole is given the opportunity to adjust the recommended passing score based on their expert judgment.

It is difficult to compare the validity of one standard setting technique to another, and this may be one reason why such a large number of techniques have been proposed. Assuming rater characteristics and training are comparable, the superiority of one method over another depends on which method produces item difficulty estimates that are most consistent with the actual performance of target examinees (Chang, 1999). However, this interpretation of validity can be highly problematic. The target examinee focused on by raters using a classical rational standard setting method is the minimally competent

candidate, while the actual population of candidates may be far beyond minimal competence. Moreover, the sample used to compute item difficulty estimates may not be representative of the entire candidate population.

In order to more accurately estimate the validity of a particular standard setting method, Chang (1999) proposed using candidates with an average (rather than minimal) performance level as a judgment reference. Chang compared the Nedelsky and Angoff, methods, showing that although intrajudge consistency was greater with the Nedelsky method, the item difficulty estimates tended to be lower than with the Angoff method (1999). These lower estimates were attributed to the discrete item difficulty estimates required for the Nedelsky (previously discussed). Although using the average candidate as a referent may be useful for the evaluation of standard setting methodology, it is impractical for applied use. If passing score studies were conducted with the average candidate used as the judges' referent, then every accurate passing score study would result in 50% of candidates passing and 50% failing when testing on a normal distribution of candidates. Smith and Smith (1988) also compared the Angoff and Nedelsky methods, demonstrating that panelists using the Angoff method used a wider variety of item information and produced item difficulty estimates that were closer to the actual *p* values of the items.

A study comparing the Yes/No Angoff and Modified-Angoff variations found that the Yes/No Angoff method produced similar passing score study results, and was considered by the panelists to be easier to understand and use (Impara & Plake, 1997). However, the Modified-Angoff method shows less potential for negative or positive bias than the Yes/No method, and has been much more widely used in both organizational and

educational contexts. When originally proposed by Angoff (1971), there was no rationale provided for either variation of this method. This omission, combined with the inability to accurately compare and evaluate different methods, may have influenced the numerous variations of this method which were subsequently proposed.

Issues in Standard Setting

Validation of a particular standard setting decision depends primarily on the assumption that the standard corresponds to the specified performance standard, in that only candidates who are minimally competent are likely to meet the standard, and those below minimal competency are not likely to pass (Kane, 1994). This is referred to as the *descriptive assumption* (Kane, 2001). Additionally, the specified standard must be reasonable given the purpose of the cut score decision. For instance, the standard for a licensure examination may reflect minimal competency, while the standard for a job knowledge test in a high-stakes selection setting may reflect a higher standard. This is referred to by Kane (2001) as the *policy assumption.* Particularly in a selection setting, assessment cut scores may serve as a reflection not only of expert rater judgment, but the testing context itself. For example, a job knowledge test used at the beginning of a multiple hurdle selection system for a job with a high selection ratio (e.g., number of jobs relative to number of applicants) may have a more lenient standard than a test used as the primary selection tool for a job with a low selection ratio.

The process of validating a performance standard involves demonstrating that the inferences and assumptions leading from the test score to the conclusions are plausible. Specifically, it must be shown that the conclusions or decisions follow from the assumptions and that the assumptions are reasonable (if a priori) or supported by the data

(Kane, 2001). According to Kane (1994), there are three types of validity evidence for evaluating performance standard decisions: internal, external, and procedural. The internal validity of the proposed interpretation of a passing score can be evaluated by empirically examining the consistency of different sets of results derived from the same passing score study (Kane, 1994). While consistency in standard setting results does not provide conclusive evidence of internal validity, it supports the assumption that the passing score is reflective of the performance standard. Hambleton and Pitoniak (2006) also note that internal validity evidence can be found by examining the consistency of the performance standard if the same techniques were replicated several times. Additional internal validity evidence is provided with intrapanelist consistency, and interpanelist (within steps) consistency. Intrapanelist consistency is a calculation of the degree to which panelist ratings vary across steps, or rounds. It is expected that the panelists taking into account the empirical data and consensus discussions associated with each round, some intrarater variability should be seen. Thus, high consistency of panelists between rounds (between-round reliability) could be evidence of poor internal validity. Interpanelist consistency is an estimation of the variability between raters within each round. High consistency between raters provides evidence of the internal validity of the performance standard. Intrapanelist and interpanelists inconsistencies may be caused by panelists having different perceptions of item or task difficulty than is empirically indicated. This phenomenon is known as disordinality (Pant, Rupp, Tiffin-Richards, & Koller, 2009).

External validity (generalizability) is evidenced through comparisons of cut scores to external sources of information (Kane, 1994). To thoroughly provide external

validity evidence, comparable standard setting results must be demonstrated among different standard setting methods and different sources of information. Additionally, the distribution of achievement which results from the application of the standard must be reasonable (Hambleton & Pitoniak, 2006). As discussed above, the stringency of performance standards is directly illustrated by the number of candidates placed in the pass and fail categories, and the reasonableness of this distribution should reflect the norms of the testing context (licensure, selection, educational, etc.). However, it is difficult to compare the performance standard with the resulting performance of candidates during administration, because the competence of each candidate cannot be fully known. Procedural validity evidence is provided by demonstrating the appropriateness of the standard setting procedure, and a high quality of implementation (Kane, 1994).

Procedural evidence is a widely accepted form of validity for policy decisions. Performance standard decisions are not deemed arbitrary if they have been arrived at by consensus with a group of individuals knowledgeable about the test content, and who understand the purpose for which the standards are being set. These content experts must also be considered unbiased, and must have a clear understanding of the standard setting process in use. This type of validity evidence is often adequate support for a particular performance standard. One argument for the importance of procedural evidence is the relative lack of other methods of validation (Pant, Rupp, Tiffin-Richards, & Koller, 2009). A second argument is that procedural evidence is widely used to validate policy decisions (Sireci, 2007). However, procedural evidence may be more useful for identifying inappropriately determined performance standard than in supporting the use

of the standards. The value of procedural validity evidence has certainly been contested by some (Haertel & Lorie, 2004; McGinty, 2005), and it is clear that to maintain the defensibility of performance standards, a variety of validity evidence must be considered. Although Kane's validity model focuses primarily on the reliability (consistency) of the standard setting process rather than the actual validity of the standard, the idea of an ideal pattern of evidence is a useful concept that can be applied to all three types of validity.

Panelist Qualifications and Study Design

According to Kane (2001), there are five main components of a test-centered standard setting procedure that influence the plausibility of the performance standard: (1) the way in which standard setting goals are defined, (2) the selection of panelists, (3) the training of panelists, (4) the definition of the performance standard, and (5) data collection procedures.

If the testing context requires a passing score to be used, it is critical to consider the primary objectives for making those pass/fail decisions. The four primary purposes for standard setting include exemplification, accountability, certification of achievement, and exhortation (Linn, 1994). Standard setting studies oriented towards exemplification would likely focus on providing concrete examples of the competencies rooted in the standards, while a test oriented towards exhortation might focus on describing the low competency currently demonstrated in a particular field. A licensure test might focus on emphasizing accountability. Thus, if the testing entity were to adopt the policy that it is less harmful to fail a truly competent candidate than to pass an unsafe practitioner, the result of this emphasis might be an increased likelihood of false-negative decisions. A test developed for credentialing or selection purposes would most likely be for a

certification of achievement. The intended purpose of determining a performance standard should be clearly stated, and should be specific enough so as to guide the panelists throughout the rest of the process (Kane, 2001).

Qualified panelists are a critical component of the validity of any performance standard, as the selection of such panelists influences not only the resulting performance standard, but perceptions of representativeness and credibility (Messick, 1995). The decisions to be made using the standard primarily determine the specific qualifications required of each panelist (Jaeger, 1991). However, standard setting panelists should generally have fairly advanced technical expertise in the related field, and should be familiar with the target population of candidates. This helps to ensure that ratings are as accurate as possible, that the verbal description of the performance standard being developed is accurate, and that the resulting standard is realistic. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) recommend that "a sufficiently large and representative group of judges should be involved to provide reasonable assurance that results would not vary greatly if the process was repeated". The *Standards* also note that the process for selecting panelists should be well documented and detailed fully. Although panelists are often selected for their expertise in a given field, a very advanced level of experience can come at the cost of representativeness (Cizek, 2006). If an entire panel of subject matter experts is more experienced than the average candidate, than this could cause ratings to be positively skewed for rational standard setting methods. However, a panel made up of inexperienced raters is likely to be less educated about test content. Both technical factors and policy issues will determine the ideal number of panelists to be used, and recommendations have varied

from as few as 5 panelists (Livingston & Zieky, 1982) to as many as 25 (Mehrens & Popham, 1992).

For qualified panelists to arrive at a valid performance standard, they should receive thorough training on the procedure and goals of the standard setting process (Mills, Melican, & Ahluwalia, 1991; Norcini, Lipner, Langdon, & Strecker, 1987). As with the selection of standard-setting panelists, training should be guided by the context and purpose of the examination. Training is often incorporated into the operational portions of a particular standard-setting method. For example, participants are often provided with feedback on their initial ratings, followed by a group discussion, before being asked to rate the items a second time. Assuming the information is intended to highlight or modify rater error, feedback can be considered one aspect of the training process (Raymond & Reid, 2001).

There are three primary tasks associated with the Modified Angoff standard-setting method: (1) acquiring an understanding of the standard-setting purpose and context, (2) developing the definition of a minimally competent candidate, and (3) estimating the proportion of minimally competent candidates who would respond to each item correctly. To train panelists' understanding of the standard-setting context, the facilitator could compare the purpose of the examination to other potential purposes, and explain test development and item writing procedures. To provide a frame-of-reference for a minimally competent candidate, the facilitator could review the educational preparation of examinees, and describe the cognitive characteristics of the target candidate. Additionally, the facilitator could provide the panelists with past examination statistics and provide examples of various proficiency levels (Raymond & Reid, 2001).

To train panelists for providing ratings, the facilitator could provide information on factors that influence item difficulty and discrimination, propose pacing and related strategies, or explain the impact of the number of response options on guessing probabilities (Raymond & Reid, 2001).

Appropriate criterion measures for well-trained participants include stability over rating occasions, consistency with assumptions underlying the standard-setting method, and ratings that reflect realistic expectations (Raymond & Reid, 2001). Although ratings tend to vary significantly for some participants across trials, Norcini and Shea (1992) demonstrated that Modified Angoff ratings aggregated over all participants are quite stable. Clauser, Swanson, and Erik (2002) also conducted a study on the impact of rater training and feedback on rater estimation error. Although the study showed that training did little to improve the raters' ability to rank order items by difficulty, there was a significant improvement in inter-judge consistency of ratings. Fehrmann, Woehr, and Arthur (1991) demonstrated similar results in a study of the effect of frame-of-reference training for raters using the Angoff procedure.

A common assumption of the Angoff standard-setting method is that participants are able to estimate the difficulty of items for minimally competent candidates. However, Impara and Plake (1998) tested this assumption and found that the panelists' ratings did not reflect this assumption. In order to further test this assumption (or identify needs for further training), Van der Linden (1982) proposed using IRT to identify inconsistencies between ratings and true item difficulties. Although defining realistic expectations for passing score study ratings is itself a subjective process, Ferdous and Plake (2005b) have shown that training is critical for demonstrating comparable judgment criteria and

cognitive processes. This study showed that differences in standard-setting ratings are partially due to the use of different decision criteria. Moreover, Fehrmann et al. (1991) showed that frame-of-reference training can lead to more accurate decisions than a non-frame-of-reference condition.

Prior to determining the passing score for an examination, the performance standard must be clearly defined. Performance levels are categories in which candidates will be classified based on examination performance, such as *Pass* or *Fail; Basic, Proficient,* or *Advanced,* etc.. Typically, policymakers and organizational stakeholders determine the performance levels to be used (Zieky, Perie, & Livingston, 2008). As with the other steps of the standard setting process, definitions of performance level should be based upon the intended use of the resulting standard. For example, although a multiple hurdles selection process might define performance levels at *Basic, Proficient,* and *Advanced,* a standard intended for licensure purposes might define a performance standard at *readiness for safe and effective practice*. Mills and Jaeger (1998) were the first to provide step-by-step recommendations for developing performance level classifications. In the Angoff procedure, panelists use preliminary performance level definitions to estimate how well candidates meeting that performance level are likely to perform on each item. As the panelists evaluate each item and are given feedback about actual candidate performance, they should be permitted to revise or expand the preliminary standards (Kane, 1994). As the number of performance levels increases, it becomes increasingly difficult to measure meaningful differences across levels. Thus, Zieky, Perie, and Livingston (2008) recommend using as few performance levels as is necessary for the purpose of the examination. Additionally, sufficient training intended to

provide a unified frame-of-reference can help to ensure that all panelists agree on a clearly stated definition of the performance standards.

The final component of a passing score study that influences the standard's plausibility is the method in which data are collected. Numerous procedures have been developed to improve the quality of data collection, and a few of these procedures have been widely adopted for use with the Modified Angoff method. If data are available on the consequences of raters' decisions, it should be provided as part of participant feedback (Busch & Jaeger, 1990; Norcini, Shea, & Kanya, 1988). Several contributors of the standard setting literature have highlighted the importance of providing normative information about typical candidate performance to prevent unreasonable standards from being recommended (Hambleton & Eignor, 1980; Jaeger, 1989; Shepard, 1980). Additionally, the use of iterative procedures, which allow panelists to review their decisions (and potentially revise them) before making a final standard determination, has been widely recommended (Busch & Jaeger, 1990; Jaeger, 1989; Shepard, 1980). Although there are potential problems associated with group dynamics (Fitzpatrick, 1989), studies have consistently shown that iterative procedures help lead panelists to more realistic performance standards (Hambleton & Powell, 1983; Jaeger, 1989; Shepard, 1980). As part of the feedback process, panelists can also receive information about how their ratings compare with the ratings of other panelists.

To provide further evidence of procedural validity, information can be collected from the panelists about their perceptions of the standard setting process, and their level of satisfaction with the process as a whole. As with other procedural validity evidence, positive evaluations do not prove that a particular method is appropriate for a given

context. A negative evaluation, however, provides cause for serious doubt of the method's usefulness. Lastly, data should be given to organizational stakeholders or other authorities for consideration of potential issues, such as adverse impact and previous pass rates. These individuals of authority may modify or reject a passing score if it seems unreasonable (Mehrens, 1986). However, any modifications must not be arbitrary and should be linked with organizational policies or actual examination data (Geisinger, 1991).

Rater Reliability

Interrater inconsistency (poor interrater reliability) is defined in the literature as a threat to the internal validity of a performance standard (Hambleton & Pitoniak, 2006; Jaeger, 1988; van der Linden, 1982). Typically, the standard deviation across individual panelist's ratings is used as an index of interrater reliability. The lower the standard deviation across panelists, the greater the interrater reliability. Organizational stakeholders are often interested in the standard deviation of aggregated Angoff ratings across panelists, because it provides information on the level of agreement on the recommended standard. It is easier to defend a particular standard when a high level of convergence has been demonstrated (Hambleton & Pitoniak, 2006). Although low convergence between panelists is typically viewed as a negative indication of the standard's internal validity, Cizek (1996) noted that individual rater differences could be an indication that a desired level of diversity has been achieved.

Berk (1996) noted several factors that can cause interpanelist consistency to drop, including the types of items being reviewed, differences in panelist backgrounds, and a lack of clarity in performance standard definitions. If the panelists used for the standard

setting meeting are not independent and unbiased, the political agenda of the panelists can even influence the consistency across raters. Jaeger (1988) examined the effectiveness of using the modified caution index with the Jaeger standard-setting method to identify aberrant patterns of panelist ratings compared to the overall group. The conclusion of this study, however, indicated that there is no clear and consistent cause of aberrant rating behavior. Plake and Impara (2001) also examined the consistency with which raters could estimate the item-level performance of minimally competent candidates. The authors compared panelists' predictions of item-level performance for minimally competent candidates to actual item-level performance of candidates scoring close to the overall performance standard in order to obtain the most reasonable comparison group to evaluate rating accuracy. As previously discussed, comparisons of Angoff ratings to actual item performance is problematic, as the actual candidate population may differ widely from the "minimally competent candidate" referent used for Angoff estimations. Their research found an average difference of only -.002, with a standard deviation of only .09. Thus, a high level of convergence can be reached, although rater reliability may be somewhat dependent on the number of panelists participating in the standard setting study. Chang (1999) showed that use of the Nedelsky method led to lower interrater reliability than use of the Angoff method for standard setting. This difference was attributed to the need for making multiple decisions per item with the Nedelsky method, and the increased focus on each response option. Thus, Chang (1999) showed that the level of rater convergence is due, in part, to the type of information judges are asked to evaluate.

Because the process of standard setting does (by necessity) have subjective elements, it has been criticized as being arbitrary (Glass, 1978). Popham (1978) countered Glass's (1978) accusation by arguing that the term 'arbitrary' can not only be defined as "capricious", but also as "involving judgment". Popham noted that the latter, however, does not necessarily imply the former, stating that "it is patently incorrect to equate human judgment with arbitrariness in this negative sense". However, item difficulty estimates within a variety of studies have been inconsistent, at times even contradictory (Bejar, 1983; Mills & Melican, 1988; Reid, 1991). Engelhard and Cramer (1992) found that the failure to accurately estimate the difficulty of an item is directly related to panelists' inconsistent ratings, which raised concern about the ability of panelists to objectively predict minimally competent candidate performance. Other studies, such as that conducted by Plake and Impara (2001) have found a high level of interrater convergence. A review by Brandon (2004) showed that, overall, Angoff estimates correlate moderately with actual item difficulty. Although it is impossible to avoid using judgment in determining a performance standard, the goal of standard-setting groups should be to make these judgments as informed as possible (Hambleton & Pitoniak, 2006). If panelists are required to take the test (without a key) prior to providing ratings, this may improve the ability of panelists to correctly estimate item difficulty.

Subjectivity has been of particular concern with the Angoff and Modified Angoff methods, as the ability of panelists to predict item-level performance of minimally competent candidates has been questioned. Although panelists are instructed to use the referent of a minimally competent candidate, raters often make Angoff estimations with the average examinee in mind. For example, a study conducted by Giraud, Impara, and

Plake (2005) found that teachers had a similar perception of minimally competent students even in different grade levels, school districts, and workshops, and with different examination content. This finding provides evidence that a common perception of minimal competency was influencing the panelists' frame-of-reference from outside of the standard setting context.

As both Glass (1978) and Hambleton and Powell (1983) have observed, panelists find it difficult to accurately predict item-level performance of candidates, even with highly reliable items. Jaeger (1982) found that the panelist's background influences the consistency of Angoff ratings in an educational setting. Plake, Melican, and Mills (1991) found that this influence was specific to the relationship between the panelist's background and the specific content of the examination. Research by Impara and Plake (1997) demonstrated that standard setting panelists more readily comprehended the rating task when using an individual (single) referent of a minimally competent candidate, rather than thinking about a group of minimally competent examinees. Although panelist predictions of the item-level performance of minimally competent candidates are certainly subjective, a high level of agreement between raters lends support to the decision-making process as being anything but arbitrary.

Group Dynamics

The majority of standard setting methods require panelists to rate items independently from other panelists. However, many methods, including the Angoff and its modifications, recommend that panelists engage in a group consensus discussion, through which the performance standard can become influenced by group dynamics and social interaction. The group consensus discussion can take several sources of

information into consideration, including organizational needs, acceptable pass/fail rates, adverse impact potential, opportunities for retesting, and the relative costs of misclassification (Geisinger, 1991). Standard setting panelists tend to appreciate the opportunity to discuss their ratings with other panelists and to receive feedback data, and these activities are typically perceived as valuable (Hambleton & Pitoniak, 2006).

One of the most popular options for adjusting a cut score via group consensus is to lower it by a standard error of measurement. Jaeger (1991) also developed an index which Geisinger (1991) termed the standard error of the test standard, which takes into account both unreliability of the test and sampling error of panelists. The performance standard can also be raised using the same methods, but a raised cut score leaves the organizational stakeholders at increased risk for legal disputes. It can be argued that changing the cut score following the systematic rating process will result in a more arbitrary procedure. However, the consensus discussion allows for adjustments to be made when there are anomalies present in the rating process, including instances when some of the panelists were not qualified, had a personal stake in the study's outcome, did not make independent ratings, or did not clearly understand the rating task (Geisinger, 1991). The *Standards* (AERA, APA, & NCME, 1999) require, however, that adjustments to the performance standard be made systematically and explicitly, with clear documentation of the procedures and rationale leading to the adjustment decision.

Nonetheless, the influence of group dynamics and the social interaction required of panelists is still of some concern, and it is unclear whether group interaction successfully leads to a more reliable standard. Studies examining the influence of group interaction on decision making has commonly found that when group members initially

(privately) favor one side of an issue over another, they will arrive at a consensus decision which is more extreme than the average of their pre-discussion positions (Fitzpatrick, 1989). These studies have stemmed from Stoner's (1961) provocative report that after discussing an issue, resulting group consensus decisions were riskier than the individual members' pre-discussion decisions. This phenomenon was later termed *group-induced polarization* by Myers and Lamm (1976). Group polarization is a robust finding which has been observed in a variety of contexts, and there is reasonable cause for concern that group polarization may also occur in standard setting (Fitzpatrick, 1989). It is unclear, however, whether group polarization in a standard setting context would be desirable or not. For example, the group consensus discussion allows panelists to be influenced by interpersonal comparisons, by cognitive learning through the information exchange, or by both processes (Fitzpatrick, 1989). However, panelists may feel pressured to alter their ratings as a result of social comparison (McGinty, 2005), and a likely result of this pressure is for the panelists with the minority position to concede to the majority position. However, a decision based on the average of both majority and minority judgment may lead to the most valid performance standard. Hertz and Chin (2002) have argued that the ideal standard setting procedure should minimize group effects and keep the process as simplified as possible.

Although social influence may make the process more subjective, it is often implied that a variety of information sources should be used to inform judges' decisions. Research conducted by Lamm and Myer (1978) has suggested that group polarization is more likely to occur when judging subjective (rather than objective) items. This evidence supports the idea that providing normative information about item or test performance

may make raters less likely to rely on social comparisons. Research has shown that providing normative information results in minor, inconsistent changes to the resulting performance standards. However, interrater reliability is increased (Busch & Jaeger, 1990). In general, group discussions have led to increased interrater consistency, which is considered to be an indication of the internal validity of a performance standard (Behuniak, Archambault, & Gable, 1982; Hurtz & Auerbach, 2003). Although valid standards must be reliable, a reliable standard does necessarily indicate high validity. Convergence amongst panelists may, in fact, be artificially derived from undesirable influences (McGinty, 2005). Having a diverse group of panelists can help to guard against group polarization or regression to the mean, as disagreement is more likely to occur when a variety of backgrounds is represented (Messick, 1995). Variability resulting from consistently different panelist values (rather than random variability around the consensus) is not necessarily undesirable (Messick, 1995).

Fatigue

Standard setting methods that require panelists to make judgments about a hypothetical candidate can be cognitively taxing for raters. This is particularly true of the Modified Angoff method, which often requires panelists to rate items over several rounds before coming to a group consensus. The amount of cognitive processing required of raters has been one criticism of the Angoff method (Lewis et al., 1998). Although a single round of Angoff ratings can be used in order to decrease panelist fatigue, the second set of estimates is considered to be more defensible, and less reflective of rater error (Ferdous & Plake, 2005a). The cognitive load placed on raters is further increased when the length of examination forms is long, or a standard setting study combines multiple

test forms in one rating session. Moreover, the requirement of a variety of professionals to serve as panelists and facilitators for a standard setting study can make the process expensive as well as time-consuming. Subject matter experts used as panelists are often active professionals in their field, and may have to miss work opportunities to participate. This costs the organizations within which they are employed as well as the panelists themselves.

Reducing the total amount of time required of panelists in a Modified Angoff standard setting study can result in lower expenses to the examination stakeholders, panelists, and panelist employers. One method to achieve this is to conduct a web-based passing score study so as to offset the costs and time of travel. Harvey and Way (1999) developed such a web-based standard-setting system, finding that the passing scores determined from a web-based study were similar to those from traditional, monitored meeting. A decrease in panelist fatigue also has the potential to increase the quality of Angoff estimates, as each panelist is left with more time and cognitive resources to review each item. If a passing score could be determined using a subset of test items (rather than the full-length test form), these issues would be alleviated.

There are two main techniques for generalizing performance standards (Ferdous & Plake, 2005a). The first is to divide test items into smaller subsets, and assign equally sized subgroups of panelists to review these subsets. This would not reduce the total number of items rated, but would reduce the total number of items rated by each panelist. However, this technique would require a large number of raters to achieve the same level of reliability and representativeness as the original modified Angoff method. The second technique involves creating a representative subset of items from the full-length test.

Standard setting panelists can then rate the smaller group of items, which should mirror the full-length test in content and psychometric properties.

Generalizing Performance Standards

Few studies have been conducted concerning the generalizability of performance standards. The National Assessment of Educational Progress (NAEP) used the panelist subgroup technique for NAEP achievement levels-setting studies, requiring each panelist within two equally sized subgroups to rate about half of the items on the examination (National Assessment Governing Board, 1994). A study conducted by Plake and Impara (2001) employed a similar strategy, splitting a 230-item test into two subsets of 115 items, based on matching psychometric properties. Each set of 115 was rated by one of two subgroups, using the traditional Angoff standard-setting procedure. Using panelist subgroups to rate subsets of test items, a group of anchor (common) items can be included on both subsets in order to verify the consistency of item ratings across groups. Results from the Plake and Impara (2001) study showed that ratings from the two subgroups were comparable when evaluating the consistency of ratings groups on 70 anchor items.

The first study to examine the generalizability of a subset of test items to the full-length exam using a single group of panelists was conducted by Sireci, Patelis, Rizavi, Dillingham, and Rodriguez (2000). The study compared three Angoff methods (including the traditional Modified Angoff method) used to set standards for Computer Adaptive Testing (CAT) items. Generalizability in standard setting is particularly important for CAT applications. The available set of items is far greater than a specified set of examination form items, and panelists cannot be expected to provide Angoff estimates on

an entire bank of items. Thus, the authors also examined the rating consistency across multiple item subsets, which each represented a third of the total item set. The study showed that only 2/3 of the item subsets were required to obtain a passing score "relatively similar" to the total 112 items (Sireci et al., 2000, p. 24). In this study, the maximum value of the deviations from the full-length test standard was one tenth of a standard deviation. Using the Angoff method, the deviations were 2.06 and 2.49 score points from the full-length test standard. One limitation of this study was the use of only a single examination within a single (educational) context, and the use of only a single group of panelists. Nonetheless, the study provided promising evidence that performance standards can be estimated with only partial items sets.

Ferdous and Plake (2005) extended the research of Sireci et al. (2000) by using stratified item sampling to create item subsets for a single group of panelists. Stratified item sampling uses content weighting and the psychometric properties of items to match each sample (subset) of items to the full-length test. This study used data from two standard setting studies for certification exams. The first standard setting study matched the subsets to the full-length test on item difficulty only. The second study matched the item subsets to the longer test on both item difficulty and content weightings. Eight item subsets were extracted from each of the full-length tests, representing between 5% and 70% of the total number of items. The performance standards determined from samples made up of half the full test length were consistently within one point of the full-length performance standard. Moreover, mean intrajudge consistency (reliability) estimates for the 50% samples were almost identical to the full-length test estimates. These results suggested that a stratified item sample of roughly 50% of the full-length examination

"may be sufficient" to estimate an equivalent passing score using the Angoff standard-setting method (Ferdous & Plake, 2005). Results were sensitive not only to the size of the sample drawn, but to the stratification procedure.

Ferdous and Plake (2007) furthered the concept of stratified item sampling by matching item subsets to the full-length test on item discrimination information, in addition to item difficulty and content weights. Modified Angoff passing score studies were conducted for elementary level Reading and Mathematics tests using fairly large samples of panelists (18 and 16 panelists, respectively). Subsets composed of 10%, 20%, 30%, 40%, and 50% of the full-length test were matched on exam content and psychometric properties. Results of this study suggested that when items are matched on content weights, item difficulty, *and* item discrimination, only 40% to 50% of items are needed to obtain a cut score within one standard error of estimate of the full-length examination cut score. Samples composed of 50% of the full test did not result in any differential impact (classifying testee as below Proficient for the sample and Proficient for the full test), and the samples composed of 40% differentially impacted no more than 1% of students (Ferdous & Plake, 2007). One limitation of this study is that the item selection technique was examined only in an educational context, with a fairly large number of raters. It is possible that stratified item selection is already being used in practice. However, more research must be conducted in order to generalize the stratified item sampling technique to other testing applications, test lengths, and panelist group sizes.

The primary purpose of this study was to evaluate the effectiveness of stratified item sampling in order to reduce the number of items needed in Modified Angoff

standard setting studies. Specifically, subsets of items were extracted from the total item set based upon content weights, item difficulty, and item discrimination. If the total amount of time required of panelists in a Modified Angoff standard setting study can be reduced while obtaining comparable performance standards, valuable financial and cognitive resources could be preserved. In order to obtain generalizable results, this study used standard setting data from a variety of industries.

Test and standard-setting characteristics varied widely across the data used for the current study. Specifically, the data sets varied widely in terms of test length, panelist group size, and interrater consistency. In addition to examining the overall feasibility of stratified item sampling, this study explored the effect of test length, panelist group size, and interrater reliability on the utility of stratified item sampling in standard setting. The potential moderating effects of these variables has yet to be examined, and it is possible that a smaller proportion of sampled items may be used when the total population of items is larger, or there are a greater number of raters providing Angoff estimates. A standard setting panel with high interrater reliability may also increase the generalizability of the resulting performance standard.

Hypotheses

1a) For theoretical stratified item subsets containing a proportion of at least 50% of the full-length test, the estimated performance standard (corrected for both finite population and the combined variance in Modified Angoff ratings accounted for by item difficulty, item discrimination, and content weighting) will be within one *standard error of the estimate* (*SEE*) of the full-length test.

1b) For theoretical stratified item subsets containing a proportion of at least 70% of the full-length test, the estimated performance standard (corrected for both finite population and the combined variance in Modified Angoff ratings accounted for by item difficulty, item discrimination, and content weighting) will be within one percentage point of the full-length test.

2a) Stratified item samplings containing a proportion of at least 50% of the total number of test items will result in a determined performance standard that is within one *SEE* of the full-length test standard in at least 95% of samples. Ferdous and Plake (2007) found that stratified item samples using 50% of the full-length test's items were well within one *SEE* of the full-length test standard.

2b) Stratified item samplings containing a proportion of at least 70% of the total number of test items will result in a determined performance standard that is within one percentage point of the full-length test standard in at least 95% of samples.

Method

Ethical and Legal Sensitivity

The researcher for this study is a current employee of the private organization from which secondary data were obtained. Ten of the 30 passing score studies used as secondary data in this study were facilitated by the researcher. However, the researcher followed standard procedure for the Modified Angoff standard setting technique, and no additional information was collected for the purpose of this study. The organization providing secondary data has asked that both the organization and the clients for which the passing score studies have been conducted be kept anonymous. Thus, detailed sample information for secondary data is excluded in order to maintain ethical and legal sensitivity.

Standard Setting Data

Standard setting data were collected from 30 independent standard setting studies in a field setting. The test and standard setting data were obtained from a variety of industries seeking licensure and certification. Testing the hypotheses using several sampling replications across various industries allows for the results of Ferdous and Plake (2007) to be more broadly generalized. See Table 1 for a summary of full-length test characteristics. Industries represented within the 30 data sets include medicine/healthcare (40%), health and safety (26.7%), technology (13.3%), security (10%), finance (3%), publishing (3%), and construction (3%). Tests varied in length from a total of 60 items to a total of 350 items (M = 134.7, SD = 50.13), and panelist group size ranged from 5

raters to 11 raters (M = 8.27, SD = 1.72). Content complexity also varied from tests containing 3 general content areas to tests containing 14 content areas (M = 5.47, SD = 1.89). While some of the tests' content areas held fairly equal weighting, others varied significantly from one content area to another.

The mean of Angoff ratings for each test is representative of the recommended minimum passing score (MPS), and this value is expressed as the percentage of items correct. Full-length MPS values across the 30 tests ranged from 65.7% correct to 79.3% correct (M = 71.5%, SD = 4.15). The most common indicator of rater reliability is the standard deviation of panelists' average Angoff ratings (Hambleton & Pitoniak, 2006). The standard deviation of mean panelist Angoff ratings varied significantly, ranging from 3.11 to 10.63 (M = 6.15, SD = 2.67). The standard error of estimate (*SEE*) was computed for each full-length test using the standard deviation of panelists' MPS values divided by the square root of panelist sample size. This mathematical computation of the *SEE* is consistent with research conducted by Ferdous and Plake (2005a; 2007). The *SEE* is often used to define the error band around the MPS, and adjustments are often allowed within one *SEE* of the original MPS. The MPS values used for this study do not include any adjustments made during the standard setting meeting. *SEE* values across the 30 tests ranged from 0.79 (10 panelists, SD = 2.09) to 4.34 (6 panelists, SD = 10.63). The average *SEE* was 2.19 (SD = 1.04).

All test form items included in standard setting were used as active items on the test form, and the psychometric property estimates (difficulty and discrimination) of each item were obtained. Item difficulty was assessed using classical test theory (p = proportion of candidates responding correctly to the item), and discrimination was

assessed using the point-biserial correlation between performance on the item and the full-length test (r_{pb}). The point-biserial is a more appropriate discrimination index than the biserial correlation when the continuous variable is not normally distributed. Item performance is treated as a dichotomous variable (pass/fail), while test performance (overall score) is treated as a continuous variable. The candidate sample size used to conduct item analyses varied, with a minimum sample size of 30 candidates required. Item statistics were collected under motivated, high-stakes conditions. Although item statistics were sometimes collected through pilot testing of items, candidates were unaware of which items were to be used as scored items on the examination. Table 2 provides the mean, standard deviation, range, skewness and kurtosis values for item difficulty and item discrimination within each set of data.

Procedure

All secondary standard setting data used for this study utilized the Modified Angoff standard setting method. This method requires subject matter experts (panelists) to individually review each item on an examination form, and to provide estimates of the proportion of minimally competent candidates who would correctly respond to the item (25%-95%). Candidates were asked to use a rating floor of 25% due to the probability of guessing correctly on a four-option multiple choice question. A 95% rating ceiling was used to account for random response error. Prior to this rating task, the definition of the target candidate and the minimally competent (or minimally acceptable) candidate is clarified for the group of panelists. Proportion estimates are then averaged across all items on the examination to compute each panelist's performance standard, and these values are averaged across raters to obtain an overall passing score. The facilitators of the

standard setting studies included in this research provided feedback in the form of a group discussion of individual panelists' ratings (average Angoff rating for each item and average overall passing score determined from ratings). However, normative information about actual item performance (difficulty or discrimination) was not shared. Following feedback from the facilitator, participants had the option to readjust their Angoff ratings. Many standard setting studies using the Modified Angoff technique end with a group consensus discussion, in which the group as a whole is given the opportunity to adjust the passing score based on their expert judgment. However, the data used as the full-length test standard (MPS) did not include any adjustments made as a result of a group consensus discussion. Instead, only the average Angoff estimates originally obtained from the panelists were compared to those obtained from the stratified item samples.

Theoretical standard estimates. Data from each standard setting study was used to compute the grand mean of all Angoff ratings, the standard deviation of panelists' ratings, the mean variance of ratings across items, and the *SEE*. Using the correction for finite populations (Scheaffer, Mendenhall, & Ott, 1979), a 95% confidence interval for the obtained cut score of each subset was calculated. The formula for the correction of finite populations is as follows:

$$\frac{\text{Variance of Angoff ratings} * (1 - \text{Adj R2 of strata})}{\text{Number of subset items}} * \text{Proportion of full length test}$$

The confidence interval ranges were corrected for finite population only, in addition to a combined correction for both finite population and the variance accounted for by subset strata. This provided an estimation of the added benefit of using the stratified item sampling procedure rather than simple random sampling from the full-length test.

Stratified item sampling. A stratified item sampling procedure was used to form multiple item subsets from each full-length test. The following proportions of items were extracted to form 6 subsets per test: 20%, 35%, 50% (replicated twice), and 70% (replicated twice). Item subsets were matched to the full-length test on content (domain) weighting, item difficulty, and item discrimination estimates. Due to the need to match a specific proportion of items to each strata category across three strata, the sampling process could not be completely random.

First, test items were classified into their respective content categories. Content weights are reflective of the tests' established examination blueprints, which are determined from job/task analyses. Next, test items within each content classification were grouped into three categories based on the difficulty level of the items. Items with difficulty levels of .00 to .50 formed Category 1, items with difficulty levels of .51 to .75 formed Category 2, and items with difficulty levels of .76 or higher formed Category 3. Lastly, items were classified into three groups based on item discrimination parameter estimates. Items with discrimination values of .10 or less formed Group 1, items with discrimination values between .11 and .35 formed Group 2, and items with discrimination values of .36 and above formed Group 3.

Each item subset was assembled to contain a proportionate number of items within each content area, and consisted of the same proportion of items within each difficulty and discrimination category as the full-length test. These three variables served as strata for the four subsets containing 20%, 35%, 50%, and 70% of the full-length examination. Two additional samples were created to replicate the assembly of the 50% and 70% subsets, on which hypotheses *2a* and *2b* were based. Repeated samplings were

assembled to contain minimal overlap with corresponding 50% and 70% subsets. The sampling process began with systematic sampling, as every 5th item was selected for the 20% subsets, every 3rd item was selected for the 35% subsets, and every 2nd item was selected for the 50% and 70% samples. Items were then replaced as necessary to obtain content weighting, difficulty, and discrimination proportions equal to that of the full-length test. Average Angoff values for each item were only visible following the completion of item sampling.

Results

Theoretical Sampling Procedure

In order to obtain the estimated variance in Angoff ratings accounted for by item difficulty, item discrimination, and content weighting, multiple linear regression was used for each full-length test, with all three variables in a combined regression block. The Angoff ratings for each item served as the dependent variable for these analyses, and thus the total number of items in the full-length test served as the sample size for each regression analysis. Item difficulty and item discrimination were analyzed as continuous variables, while content weightings were dummy-coded as dichotomous variables. For example, a test with 5 content areas was dummy-coded into 4 categorical variables so that the regression model contained 4 degrees of freedom for the content weightings, one degree of freedom for item difficulty, and one degree of freedom for item discrimination. The adjusted R^2 obtained from each of these regression analyses were then plugged into the correction for finite population formula as follows:

$$\frac{\text{Variance of Angoff ratings} * (1 - \text{Adj R2 of strata})}{\text{Number of subset items}} * \text{Proportion of full length test}$$

The fully corrected formula can be compared to the formula correcting only for finite population in order to estimate the added benefit of using a stratification procedure rather than simple random sampling. Table 3 lists the variance in ratings accounted for (adjusted

R^2) by item difficulty, item discrimination, and content weightings separately as well as combined.

Hypothesis 1a stated that theoretical stratified item subsets containing a proportion of 50% of the full-length test would provide an estimated performance standard (corrected for both finite population and the combined variance in Modified Angoff ratings accounted for by item difficulty, item discrimination, and content weighting) that is within one *standard error of the estimate* (*SEE*) of the full-length test. The theoretical sampling estimates fully supported this hypothesis. When conducting theoretical simple random sampling across the 30 data sets, an estimated proportion of 40% of the full-length test was required to obtain a cut score within one *SEE* of the full-length test standard (*SD* = 0.16). This estimation accounted for rating variance across test items. On average, the theoretical stratified item samplings indicated that only 35% of the full-length test would be necessary to obtain MPSs within one *SEE* of the full-length test (*SD* = 0.17). Figure 1 shows the proportion of items required to obtain a cut score within one *SEE* of the full-length test standard, using stratified or simple random sampling. The correction for finite population used for the theoretical subsets accounted for the variance in ratings as well as the variance in ratings accounted for by subset strata.

Hypothesis 1b stated that theoretical stratified item subsets containing a proportion of at least 70% of the full-length test, the estimated (fully-corrected) performance standard will be within one percentage point of the full-length test. This hypothesis was also fully supported by the theoretical data. On average, 71% of the full-length test was required for simple random sampling in order to reliably obtain a MPS within one percentage point of the full-length test standard (*SD* = 0.12). The theoretical

stratified item sampling indicated that only 65% of the full-length test was required in order to obtain an MPS within one percentage point of the full-length test MPS (*SD* = 0.11). Table 4 provides a list of the estimated proportion of each full-length test required to reliably obtain a cut score within one *SEE* and one percentage point of the full-length test MPS. Figure 2 shows the proportion of items required to obtain a cut score within one *SEE* of the full-length test standard, using stratified or simple random sampling.

The adjusted R^2 of the Angoff ratings using item difficulty as a predictor varied from -.01 to .48, with a mean adjusted R^2 of .18 (*SD* = .14). The adjusted R^2 of item discrimination (r_{pb}) on Angoff ratings ranged from -.06 to .17 (*M* = .01, *SD* = .04). The adjusted R^2 of content weighting on Angoff ratings also varied significantly from -.02 to .36, with a mean adjusted R^2 of .24 (*SD* = .15). The adjusted R^2 for the combined set of strata ranged from -.03 to .47, with a mean of .24 (*SD* = 15). Item difficulty was a significant predictor ($p < .05$) of Angoff ratings for 26 of 30 data sets (86.7%), item discrimination significantly predicted ratings in 5 of 30 data sets (16.7%), and content weighting predicted Angoff ratings in 12 of 30 data sets (40%). The combined set of strata significantly predicted Angoff ratings in 26 of 30 data sets (86.6%). When the combined set of strata did not account for variance in the Angoff ratings, there were no differences between the estimated MPS values of the theoretical simple random samples and those of the theoretical stratified item samples. Figures 3, 4, 5, and 6 show the relationship between item difficulty and corresponding Angoff estimates for the four tests that did not show a relationship between test strata and Angoff ratings (tests 18, 22, 28, and 29).

To evaluate potential multicollinearity between item difficulty and item discrimination, the Pearson's product-moment correlation coefficient for the relationship between item difficulty and discrimination was computed for each sample. The correlation between these two variables is detailed for each sample in Table 3. The relationship between item difficulty and item discrimination varied widely across the 30 samples, ranging from -.32 to .44, with a standard deviation of .21. While the difficulty and discrimination correlation was significant for 21 of the 30 data sets, the nature of the relationship is inconsistent. However, there was a significant correlation between item difficulty and item discrimination for each of the five data sets which showed item discrimination as a significant predictor of Angoff ratings.

Applied Sampling Procedure

In addition to using a theoretical sampling procedure to identify the estimated subset proportion required to obtain a MPS comparable to the full-length test standard, this was also tested using actual stratified item subsets derived from each full-length test. Table 5 lists the absolute value of the difference between the full-length test MPS and the MPS derived from each of 6 subsets from each data set (20%, 35%, two 50%, and two 70% subsets). *Hypothesis 2a* stated that a stratified item sampling containing a proportion of 50% of the total number of test items will result in a determined performance standard that is within one *SEE* of the full-length test standard (in at least 95% of samples). To test this hypothesis, the MPSs of the two 50% subsets from each data set were compared to the MPS obtained from the full-length test. Repeated samplings conducted on all 50% subsets allowed the results of the study to be more conclusive, and reduced the aggregated effects of sampling error. *Hypothesis 2a* was fully supported. The absolute

values of the difference between the 50% subset MPSs and the full-length test MPSs were less than one *SEE* for 58 of the 60 subsets. Incidentally, the two subsets that did not meet this criteria were derived from the same full-length test, and the *SEE* corresponding to this data set was less than one (*SEE* = .79). The average absolute MPS difference value across the 60 (50%) subsets was less than 1 percentage point (*M* = 0.58%, *SD* = 0.46). To test the significance of these results, a chi-square test was conducted. Using an expected value of 95% of the 60 subsets falling within one *SEE* of the full-length test MPS, the obtained χ^2 value (1, *N* = 60) was .35, *p* = .55. Thus, there is no statistically significant difference between the expected percentage and the observed percentage.

Hypothesis 2b stated that stratified item samples containing a proportion of 70% of the full-length test will result in a determined performance standard that is within one percentage point of the full-length test standard (for at least 95% of samples).This hypothesis was also fully supported, as the difference between the full-length test MPS and the subset MPS was less than one percentage point for 59 of 60 subsets. The average absolute MPS difference value across the 60 (70% proportion) subsets was 0.37%, *SD* = 0.30. Using an expected value of 95% of the 60 subsets falling within one percentage point of the full-length test MPS, the obtained χ^2 value (1, *N* = 60) was 1.40, *p* = .24. This χ^2 value indicates that no statistically significant difference exists between the expected percentage and the observed percentage.

For each of the 30 data sets, one 20% subset and one 35% subset was also derived. These samples often produced MPS values that were comparable to the full-length test standard. Across the 20% subsets, the subset MPS was within one *SEE* of the full-length test standard in 27 of 30 samples (90%). The 20% subset MPSs were within

one percentage point of the full-length test standard in 22 of 30 samples (73.3%). Across the 35% subsets, the subset MPS was within one *SEE* of the full-length test standard in all 30 samples (100%). The 35% subset MPSs were also within one percentage point of the full-length test standard in 22 of 30 samples (73.3%). The average MPS difference value was 0.91 (*SD* = 0.76) across the 20% subsets and 0.68 (*SD* = 0.38) across the 35% subsets. These results indicate that subsets containing less than 50% of the full-length test can produce MPSs comparable to the full-length test standard, but are not likely to do so reliably.

Test and Standard Setting Characteristics

The effects of test length, panelist group size, and interrater reliability on the utility of stratified item sampling in standard setting were also examined as secondary analyses. In order to examine the potential moderating effects of these variables, linear regression was used to evaluate the influence of various characteristics of the standard setting process (test length, number of panelists, and interrater reliability) on the size of the absolute difference scores between the full-length test MPS and each 20% subset MPS. In order to avoid violating the assumption of independence, the 6 subsets derived from each full-length test were not combined. Instead, only the 20% subsets were used to test the effects of test and standard setting characteristics, as the MPS difference scores were expected to be larger than they would be if a larger proportion of items were used.

The analyses indicated that test length does not significantly predict the generalizability of stratified item subsets to corresponding full-length tests (R^2 = .01, p = .65). Similarly, panelist group size did not significantly increase or decrease the difference in MPS between 20% subsets and the full-length tests (R^2 = .01, p = .64). To

examine the effect of interrater consistency on stratified item sampling, the standard deviation across judges in each data set was used as an independent variable in linear regression to test for the effect of interrater reliability on the generalizability of subset MPSs. Across the 20% subsets (N = 30), interrater reliability was not a predictor of the size of the MPS difference scores (R^2 = .04, p = .28).

Discussion

Theoretical Sampling Procedure

It was predicted in *Hypothesis 1a* that theoretical stratified item subsets containing a proportion of at least 50% of the full-length test would provide a performance standard within one *SEE* of the full-length test. The theoretical sampling estimates fully supported this hypothesis. On average, only 35% of the full-length test was required to obtain a cut score within one *SEE* of the full-length test standard when stratified item sampling was estimated. These estimated proportions based on theoretical stratified item sampling are consistent with the required proportions cited by Ferdous and Plake (2007, 2005b). Using item difficulty, discrimination and content weightings as strata in stratified item sampling, Ferdous and Plake (2007) reported that samples of 40-50% of the full-length test were adequate to obtain standards within one *SEE* of the full-length test. In a previous study using only item difficulty and content weightings as strata, Ferdous and Plake found that approximately 50% of the full-length test items were required (2005b). In the studies conducted by Ferdous and Plake, the addition of item discrimination as a variable in the stratified item sampling process reduced the required sampling proportion by as much as 10%. However, item discrimination did not serve as a significant predictor of Angoff ratings in the majority of the data sets analyzed in this study. In the few cases when item discrimination did serve as a significant predictor of Angoff ratings, discrimination values were also significantly correlated with item difficulty values. Thus, the relationship between item discrimination values and Angoff ratings may simply have been reflective of extreme difficulty values (e.g., easy items tend

to have low discrimination values). Therefore, although the results of this study do support the use of stratified item sampling, the findings do not support the use of item discrimination as a stratification variable.

Item difficulty and content weightings accounted for a larger proportion of variance in Angoff ratings than did item discrimination. Item difficulty was a significant predictor of Angoff estimates in 26 of 30 data sets, while content weightings significantly predicted ratings in 12 of 30 data sets. The results of this study support the use of stratified item sampling with only item difficulty and content weightings serving as strata, because stratifying based on item discrimination is not likely to result in reduced sampling error over simple random sampling. The estimated proportion of items required to obtain a cut score within one *SEE* of the full-length test standard using simple random sampling was 40%. Simulations across 30 standard setting studies indicated that using a stratified item sampling procedure, rather than simple random sampling, reduced the required proportion of the full-length test by an average of 5%. However, the extent to which stratified item sampling reduces sampling error depends on the variance in Angoff ratings accounted for by the sampling strata. The variance in ratings accounted for by item difficulty, discrimination and content weightings varied widely across the 30 data sets, with estimated required proportions up to 12% less than simple random sampling estimates in some samples. In applied applications, the variance in ratings accounted for by these strata would be unknown at the time stratified item subsets were created for a standard setting study. Given that the added benefit of stratification could be significant, using a stratification procedure is always recommended.

A larger required proportion of the full-length test was expected when the criterion for a 'comparable' standard is more conservative. *Hypothesis 1b* predicted that the estimated cut scores for theoretical stratified item subsets containing a proportion of at least 70% of the full-length test will be within one percentage point of the full-length test standard. This hypothesis was also fully supported by the theoretical cut score estimates. When a stratified item sampling procedure was theoretical, an average of 65% of the full-length test was required in order to reliably obtain a MPS within one percentage point of the full-length test standard. The average required proportion simulating only simple random sampling was 71%, indicating an added benefit of 6% when using stratification. Across all 30 data sets, the decrease in the required proportion of items when the strata variables were accounted for ranged from 0% to 14%. When the criterion for a comparable standard is set conservatively at one percentage point above or below the full-length test standard, only about 2/3 of the total number of test items is required to obtain a comparable cut score 95% of the time. These results are consistent with the findings of Sireci and colleagues (2000). The theoretical sampling estimates also support the use of a stratified item sampling procedure when the criterion is set to within one *SEE* of the full-length test standard. The theoretical sampling indicated that subsets containing only 35% of the full-length test are sufficient to obtain cut scores within one *SEE* of the full-length test 95% of the time. This corresponds with the results of Ferdous and Plake (2007; 2005b).

Using the *SEE* as an indicator of cut score comparability is reasonable given that adjustments to the cut score are most commonly made using a similar index of reliability, the *standard error of measurement* (Hambleton & Pitoniak, 2006). Both indices account

for interrater reliability and the number of raters used in standard setting. However, the variability of this index makes it unreliable for use as a criterion for cut score comparability. When the number of raters and interrater reliability are low in a standard setting study (resulting in a high *SEE*), it is significantly easier to obtain a 'comparable' standard. However, when interrater reliability is high in a standard setting study with many raters, the *SEE* can be very low. This makes it difficult to reliably obtain a standard within one *SEE* of the full-length test standard because the comparability criterion becomes very conservative. Applications of the results of this research should rely on a more consistent and conservative index of comparability, given that stratified item subsets are created prior to obtaining interrater reliability.

Applied Sampling Procedure

In order to expand on the applied research of Ferdous and Plake, stratified item subsets were derived from 30 licensure and certification tests amongst a wide variety of organizations and industries. To compare the applied use of a stratified item sampling procedure to the sampling simulation, six stratified item subsets were composed from each data set (20%, 35%, two 50% subsets, and two 70% subsets). The sampling simulation procedure estimated various subset MPSs when *simple random sampling* was used, while factoring in the variance in ratings accounted for by sampling strata. In applications of stratified item sampling, however, sampling could not be completely random when stratifying occurs across three variables. While items were originally selected using systematic sampling within the content areas of a test, item substitutions were made as necessary to ensure that proportions of items within each difficulty and discrimination category were representative of the full-length test. Substitutions made

during the sampling process were not completely random, although Angoff ratings were unknown during item selection. This sampling procedure is reflective of the sampling method which would most likely be used when creating stratified item subsets for the purpose of standard setting.

Full support was obtained for *Hypothesis 2a*, which stated that a stratified item sampling containing a proportion of 50% of the total number of test items will result in a determined performance standard which is within one *SEE* of the full-length test standard (in at least 95% of samples). The observed MPS values from a total of 60 subsets (containing 50% of the full-length test items) were compared to the MPS obtained from the full-length test. The absolute values of the difference between the 50% subset MPSs and the full-length test MPSs were less than one *SEE* in 96.67% of the 60 subsets. The two subsets which did not meet this criterion were derived from the same full-length test, and the *SEE* corresponding to this data set was very low. The average absolute MPS difference value across these 60 subsets was less than 1 percentage point.

Hypothesis 2b, which stated that stratified item samples containing a proportion of 70% of the full-length test will result in a performance standard that is within one percentage point of the full-length test standard, was also fully supported. The difference between the full-length test MPS and the 70% subset MPS was less than one percentage point for 98.33% of 60 subsets. The average absolute MPS difference value across the 70% subsets was .37. Additionally, subsets containing only 20% or 35% of the full-length test often produced MPS values that were comparable to the full-length test standard. Across the 20% subsets, the subset MPS was within one *SEE* of the full-length test standard in 90% of 30 samples, and was within one percentage point of the full-length

test standard in 73.33% of the samples. Across the 35% subsets, the subset MPS was within one *SEE* of the full-length test standard in all 30 samples, and was also within one percentage point of the full-length test standard in 73.33% of samples. The average MPS difference value was 0.91 across the 20% subsets and 0.68 across the 35% subsets.

The results from the applied stratified sampling procedure indicate that subsets containing between 50% and 70% of the items on the full-length test can reliably produce MPS values comparable to the full-length test standard, depending on the comparability criterion applied. Subsets containing less than 50% of the full-length test can also produce MPSs comparable to the full-length test standard, but are not likely to do so reliably. These findings replicate the results obtained from similar research on the use of item subsets in standard setting (Ferdous & Plake, 2007; 2005b; Sireci et al., 2000). Although research conducted by Ferdous and Plake (2007, 2005b) demonstrated a marked decrease in the required proportion of the full-lenth test when using stratification(40-50%), Sireci and colleagues (2000) demonstrated that only 2/3 of the full-length test is required when simple random sampling is used. Simple random sampling was not used in this study as a means of comparison to the applied stratified item sampling procedure. However, simple random sampling (with no stratification) was theoretical for the purpose of assessing the added benefit of stratification. The results of these simulations indicate that the use of stratification (over simple random sampling) decreases the required subset proportion by an average of 5% when one *SEE* is used as the comparability criterion. When one percentage point is used as the comparability criterion, the use of stratification when creating item subsets decreases the required subset proportion by an average of 6%. The benefit of stratification is dependent on the

degree to which strata account for the variance in Angoff ratings. However, the advantage of stratification can only be known prior to sampling if item performance parameters (e.g., difficulty and discrimination) are available prior to standard setting. When item statistics for every item are available, the use of a stratified item sampling method is recommended. When item statistics are not available, stratification on content weightings may result in a more comparable cut score than when using simple random sampling, particularly if some content areas are more difficult than others. This sampling procedure has the potential to significantly reduce the total amount of time required of panelists in a Modified Angoff standard setting study, which in turn could preserve valuable financial and cognitive resources.

Test and Standard Setting Characteristics

Secondary analyses were conducted to examine the effect of test and standard setting characteristics on the generalizability of stratified item subsets, as the effect of these variables had not previously been examined in the context of standard generalizability. However, these variables did not appear to significantly influence the feasibility of stratified item sampling in standard setting. Across the 30 subsets containing 20% of the corresponding full-length tests, test length, panelist group size, and interrater reliability did not significantly predict MPS difference scores.

The smallest subsets (20%) were used to test the effects of test and standard setting characteristics because the corresponding MPS difference scores were expected to be greater (due to increased sampling error) than those of larger subsets. While the average MPS difference scores for the 20% subsets were larger than the average difference scores of larger subsets, the average magnitude of the difference scores was

still very small (less than one). Additionally, although the variance across the MPS difference scores for these subsets was also larger than the variance of difference scores of larger subsets, the variance in MPS generalizability may have been too small to observe a robust relationship between standard setting characteristics and MPS generalizability. Although additional research should be conducted on the moderating effects of these and other standard setting characteristics, stratified item sampling is even more valuable if subset standards can generalize to the full-length test standard regardless of test length, number of panelists, or interrater reliability.

Limitations and Future Directions

A few important limitations of this study should be noted. One limitation is the retrospective nature of data analysis. If an experimental design had been utilized, more information about the effects of rater fatigue in standard setting may have been obtained. For example, if panelists were randomly assigned to provide Angoff ratings for either the full-length test or a particular subset of the test, the effects of rater fatigue on Angoff ratings could be directly assessed. Without measuring the results of fatigue on Angoff ratings, it cannot be assumed that panelists would provide the same rating for an item on a smaller subset of items as they would if the item were part of the full-length test. However, the use of an experimental design would require that each panelist provide multiple sets of ratings on both the full-length test and one or more stratified item subsets. With this design, comparisons of repeated sets of ratings would not be a useful way to assess rater fatigue, as both fatigue and practice effects would still apply. Alternatively, different panelists could be randomly assigned to rater either the full-length test or a stratified item subset. However, it would be unclear whether fatigue or rater effects were

accounting for differences in Angoff ratings. The primary purpose of this study was to assess the feasibility of using stratified item sampling as a means for reducing sampling error. However, additional research should be conducted in order to assess the effects of cognitive load on the Modified Angoff standard setting process. Although the potential for rater fatigue is a major criticism of this standard setting method (Lewis et al., 1998), experimental research on the actual effects of fatigue on Angoff ratings is lacking.

Another limitation of this study was the lack of variance in subset MPS values, which made it unfeasible to analyze the potential moderating effects of standard setting and test characteristics. The results of this study suggest that these characteristics have no effect on the generalizability of stratified item subsets. However, other research on the test subset generalizability in standard setting should continue to analyze the effects of standard setting characteristics, particularly when generalizability is low.

As previously noted, the theoretical sampling procedure predicted the range of stratified item subset MPS values when simple random sampling was used, while factoring in the variance in ratings accounted for by sampling strata. In applications of stratified item sampling, however, sampling could not be completely random when stratifying occurs across three variables. This is reflective of most examination form assembly methods, in which items are often selected for inclusion in a form based on item statistics and a particular distribution of items within each content area.

In summary, this study was conducted to evaluate the effectiveness of stratified item sampling for reducing the number of items rated in Modified Angoff standard setting studies, and to determine the proportion of test items required to obtain standards comparable to the full-length test. Representative subsets of items were extracted from

the total item set based upon content weights, item difficulty, and item discrimination, and the generalizability of these subsets were compared to the subset standard estimates obtained from sampling simulations. The results of this study provide full support for the feasibility of allowing standard setting panelists to rate smaller subsets of test items rather than providing Angoff ratings for every item. This method of generalizing a standard reduces the total amount of time required of panelists during the standard setting process, which in turn can lower expenses to examination stakeholders, panelists, and panelists' employers. Panelists may also reserve valuable cognitive resources and reduce fatigue, which can increase the quality of ratings. Stratified item sampling may provide further benefits to organizations particularly concerned about test security and item exposure, or to organizations utilizing CAT methods with very large item pools.

Table 1

Test and Standard Setting Characteristics

Test	Number Items	Number Raters	Content Areas	Full Test *MPS*	*SD* of Panelists' Ratings	*SEE*
1	60	9	6	72.55	8.92	2.97
2	110	6	6	67.87	10.03	4.09
3	175	10	5	71.02	10.18	3.22
4	110	6	6	68.64	10.39	4.24
5	125	10	4	73.41	6.71	2.12
6	150	5	5	66.16	7.30	3.27
7	110	6	6	68.27	10.63	4.34
8	150	7	5	72.30	5.44	2.06
9	150	6	5	74.88	8.98	3.67
10	175	10	5	70.77	8.78	2.78
11	125	10	4	79.15	2.09	0.79
12	80	9	7	68.26	3.22	1.07
13	150	7	5	71.33	4.81	1.82
14	125	10	4	74.02	6.78	2.14
15	175	10	5	68.43	10.21	3.23
16	150	10	7	66.05	5.02	5.02
17	100	7	4	66.43	5.25	1.75
18	150	7	5	73.09	4.05	1.53
19	100	7	4	67.47	5.64	1.88
20	175	10	5	71.03	9.43	2.98
21	125	7	4	74.28	3.18	1.18
22	150	7	3	72.95	4.23	1.60
23	125	11	6	65.69	4.41	1.33
24	100	8	5	76.87	3.11	1.10
25	100	8	6	79.34	4.15	1.47
26	126	8	6	76.98	5.00	1.77
27	100	8	5	79.34	3.76	1.33
28	100	11	5	72.03	3.54	1.07
29	350	8	14	69.08	6.09	2.15
30	120	10	7	65.84	3.20	1.01
M	**134.70**	**8.27**	**5.47**	**71.45**	**8.19**	**2.30**
SD	**50.13**	**1.72**	**1.89**	**4.15**	**11.25**	**1.15**

Note. Test is passing score study data for each full-length test. *MPS* = minimum passing score. *SEE* = *SD* of panelists' ratings/square root of number of panelists.

Table 2

Item Difficulty and Discrimination Descriptive Statistics

	Item Difficulty					Item Discrimination				
Test	*M*	*SD*	*Range*	*Skew*	*Kurt*	*M*	*SD*	*Range*	*Skew*	*Kurt*
1	0.81	0.20	0.74	-1.48	1.43	0.19	0.24	1.19	-0.51	0.47
2	0.74	0.13	0.66	-0.45	0.40	0.31	0.10	0.61	-0.64	1.11
3	0.79	0.18	0.73	-0.59	-0.48	0.14	0.28	1.54	-0.17	0.27
4	0.74	0.14	0.97	-1.61	5.89	0.31	0.10	0.65	-0.42	1.32
5	0.68	0.13	0.66	-0.47	-0.08	0.38	0.15	0.74	0.39	0.39
6	0.78	0.25	0.82	-1.32	1.09	0.10	0.23	1.16	0.82	0.53
7	0.74	0.13	0.66	-0.50	0.27	0.30	0.10	0.53	-0.73	0.86
8	0.75	0.12	0.61	-0.54	0.15	0.24	0.13	0.66	-0.46	-0.14
9	0.80	0.22	0.84	-1.22	0.66	0.14	0.25	1.10	0.18	-0.52
10	0.80	0.17	0.55	-0.61	-0.68	0.14	0.26	1.41	-0.33	0.41
11	0.73	0.11	0.57	-1.13	1.41	0.37	0.15	0.74	0.45	0.09
12	0.68	0.19	0.96	-1.93	4.99	0.23	0.20	0.58	0.05	-1.53
13	0.67	0.22	0.96	-1.77	3.48	0.24	0.19	0.58	-0.10	-1.45
14	0.70	0.14	0.77	-0.87	1.31	0.37	0.15	0.74	0.45	0.09
15	0.70	0.23	0.64	-0.79	-0.17	0.16	0.31	1.54	-0.37	0.22
16	0.68	0.19	0.88	-0.83	0.39	0.08	0.05	0.47	1.15	7.40
17	0.72	0.11	0.50	0.06	-0.61	0.24	0.11	0.70	0.35	1.27
18	0.76	0.15	0.73	-0.82	0.65	0.23	0.17	1.13	-0.81	2.15
19	0.72	0.11	0.50	-0.09	-0.57	0.25	0.12	0.66	0.10	0.05
20	0.77	0.24	0.67	-1.16	0.88	0.10	0.26	1.53	0.11	0.29
21	0.68	0.17	0.80	-0.63	0.04	0.35	0.15	0.69	0.15	-0.52
22	0.76	0.14	0.71	-0.48	-0.03	0.23	0.16	0.97	-0.40	1.23
23	0.78	0.11	0.36	0.35	-1.11	0.26	0.14	0.45	0.48	-0.78
24	0.73	0.14	0.97	-2.76	12.64	0.29	0.13	0.61	-0.68	0.32
25	0.72	0.10	0.43	-0.88	0.28	0.34	0.13	0.59	0.11	-0.31
26	0.68	0.13	0.66	-0.70	-0.29	0.37	0.14	0.70	0.22	0.22
27	0.73	0.17	0.78	-1.78	6.24	0.21	0.23	1.16	-0.30	-0.59
28	0.75	0.18	0.82	-1.61	4.40	0.15	0.15	0.67	0.03	-0.41
29	0.66	0.14	0.60	-0.59	-0.67	0.38	0.14	0.69	0.34	0.31
30	0.68	0.14	0.84	-1.01	1.11	0.36	0.14	0.66	0.18	-0.38

Note. Test is passing score study data for each full-length test. Skew = skewness. Kurt = Kurtosis.

Table 3

Variance in Ratings Accounted for by Individual Strata and Combined Strata

Test	*p*	r_{pb}	Content	Combined	*p* and r_{pb} Correlation
1	0.36*	0.17*	-0.02	0.41*	0.34*
2	0.08*	0.00	0.29*	0.39*	0.44*
3	0.37*	0.02*	0.01	0.37*	0.29*
4	0.05*	0.00	0.36*	0.42*	0.41*
5	0.27*	0.06*	-0.02	0.26*	0.27*
6	0.33*	0.00	0.17*	0.46*	-0.03
7	0.04*	0.01	0.27*	0.34*	-0.06
8	0.09*	0.00	0.09*	0.18*	-0.03
9	0.25*	0.03*	0.21*	0.41*	-0.26*
10	0.27*	0.00	0.02	0.29*	-0.25*
11	0.32*	0.01	0.13*	0.36*	0.13*
12	0.10*	-0.01	0.01	0.07	0.24*
13	0.08*	0.02	0.06*	0.15*	0.10
14	0.20*	0.00	0.00	0.20*	0.17*
15	0.48*	0.01	0.00	0.47*	-0.21*
16	0.17*	-0.01	0.05*	0.20*	-0.06
17	0.24*	-0.01	-0.02	0.21*	-0.17*
18	-0.01	-0.01	0.01	-0.01	-0.12*
19	0.20*	-0.01	0.01	0.19*	-0.16*
20	0.43*	0.00	0.04*	0.44*	-0.11*
21	0.11*	0.03*	0.10*	0.18*	0.29*
22	0.01	-0.01	0.02	0.02	-0.04
23	0.07*	0.00	0.01	0.10*	-0.19*
24	0.05*	0.00	0.03	0.07*	0.19*
25	0.15*	0.02	0.04	0.19*	0.07
26	0.25*	-0.01	0.03	0.26*	0.15*
27	0.20	-0.01	0.03	0.21*	0.03
28	0.01	-0.06	-0.02	-0.03	0.13*
29	0.01*	0.00	0.05*	0.04*	-0.32*
30	0.30*	0.00	0.01	0.30*	-0.01
M	**0.18**	**0.01**	**0.07**	**0.24**	**0.04**
SD	**0.14**	**0.04**	**0.10**	**0.15**	**0.21**

Note. Test is passing score study data for each full-length test, and corresponding values represent the adjusted R^2 value for each variable or combination of variables. *$p < .05$. Separate regression analyses were conducted to estimate the independent effect of each variable on Angoff ratings. *p* = item difficulty, r_{pb} = item discrimination

Table 4

Estimated Proportion of Full-length Test Required to Obtain Comparable Standard

Test	1 *SEE*		1 %	
	Not Stratified	*Stratified*	*Not Stratified*	*Stratified*
1	32%	22%	80%	70%
2	32%	22%	89%	83%
3	21%	14%	73%	63%
4	30%	20%	89%	82%
5	32%	26%	67%	61%
6	33%	21%	85%	74%
7	29%	21%	89%	84%
8	37%	33%	71%	67%
9	26%	17%	83%	74%
10	31%	24%	77%	71%
11	54%	43%	67%	57%
12	65%	62%	67%	66%
13	41%	37%	69%	66%
14	30%	26%	66%	61%
15	23%	14%	77%	63%
16	27%	23%	49%	43%
17	57%	51%	80%	76%
18	50%	50%	70%	70%
19	42%	37%	72%	68%
20	27%	17%	76%	65%
21	71%	67%	61%	56%
22	47%	47%	69%	69%
23	50%	47%	64%	62%
24	35%	34%	40%	38%
25	28%	26%	46%	43%
26	42%	35%	70%	63%
27	52%	46%	66%	60%
28	57%	57%	60%	60%
29	23%	22%	58%	57%
30	86%	80%	86%	81%
M	**40%**	**35%**	**71%**	**65%**

Note. Test is passing score study data for each full-length test, and corresponding values represent the proportion of the full-length test required to obtain a stratified item subset cut score within one standard error of estimate (*SEE*) or one percentage point of the full-length test cut score. Stratified item subsets were created using item difficulty, item discrimination, and content weightings as strata.

Table 5

Absolute Value MPS Differences between Full-length Tests and Corresponding Subsets

Test	20%	35%	50%	50% (2)	70%	70% (2)
1	0.47	1.29	1.15	0.36	0.00	0.06
2	0.06	0.35	1.55	0.13	0.71	0.20
3	0.98	0.15	0.76	0.15	1.07	0.6
4	2.37	0.93	0.41	0.30	0.36	0.87
5	0.31	0.35	1.05	0.65	0.38	0.27
6	0.72	1.12	0.04	0.67	0.20	0.27
7	0.69	0.84	1.28	2.45	0.03	0.50
8	0.21	1.03	0.17	0.21	0.47	0.04
9	0.65	0.54	0.98	1.12	0.92	0.52
10	1.36	1.23	0.32	0.63	0.10	0.92
11	0.34	0.42	1.34	1.05	0.74	0.12
12	0.86	0.37	0.33	0.03	0.39	0.47
13	0.53	0.55	0.53	0.6	0.06	0.31
14	0.69	1.41	1.04	1.04	0.09	0.25
15	1.72	1.00	0.30	0.48	0.04	0.09
16	0.07	0.44	0.26	0.37	0.38	0.47
17	0.63	0.06	0.40	0.22	0.44	0.89
18	0.93	0.67	0.56	0.05	0.15	0.20
19	2.42	0.46	0.07	0.11	0.20	0.80
20	0.14	1.06	0.63	0.89	0.35	0.48
21	0.30	0.00	0.83	1.12	0.21	0.36
22	1.81	0.78	0.76	0.44	0.09	0.21
23	2.08	0.80	0.02	0.10	0.29	0.33
24	0.42	0.77	0.26	0.60	0.12	0.04
25	0.03	0.80	0.22	0.10	0.03	0.11
26	0.57	0.19	0.92	0.56	0.19	0.01
27	0.99	1.02	0.34	0.30	0.19	0.21
28	0.94	0.28	0.91	1.04	0.95	0.09
29	1.31	0.73	0.30	0.64	0.84	0.87
30	2.84	0.78	0.50	0.25	0.96	0.63
M	**0.91**	**0.68**	**0.61**	**0.55**	**0.37**	**0.37**
SD	**0.76**	**0.38**	**0.42**	**0.50**	**0.32**	**0.28**

Note. Test is passing score study data for each full-length test. MPS = minimum passing score. Subsets are assembled using item difficulty, item discrimination, and content area weights as strata.

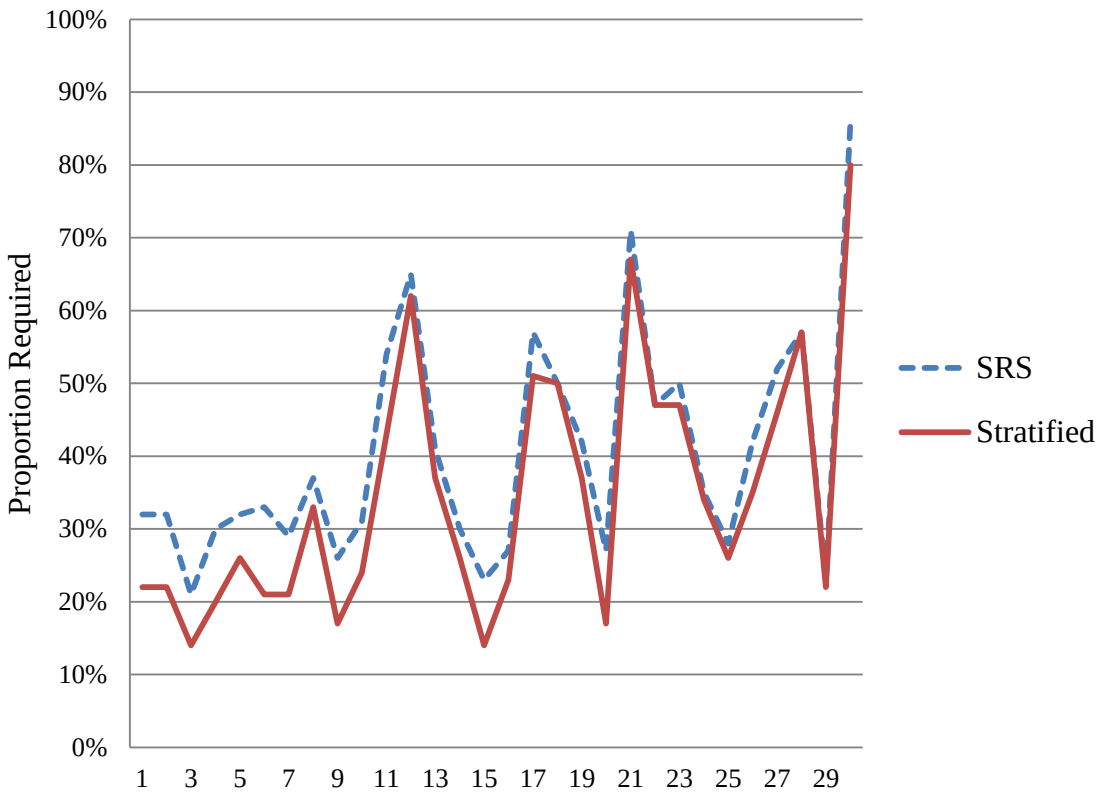

Figure 1. *Proportion of Test Required for Standard Within One SEE of the Full-Length Test.* SRS= simple random sampling. Stratified = stratified item sampling

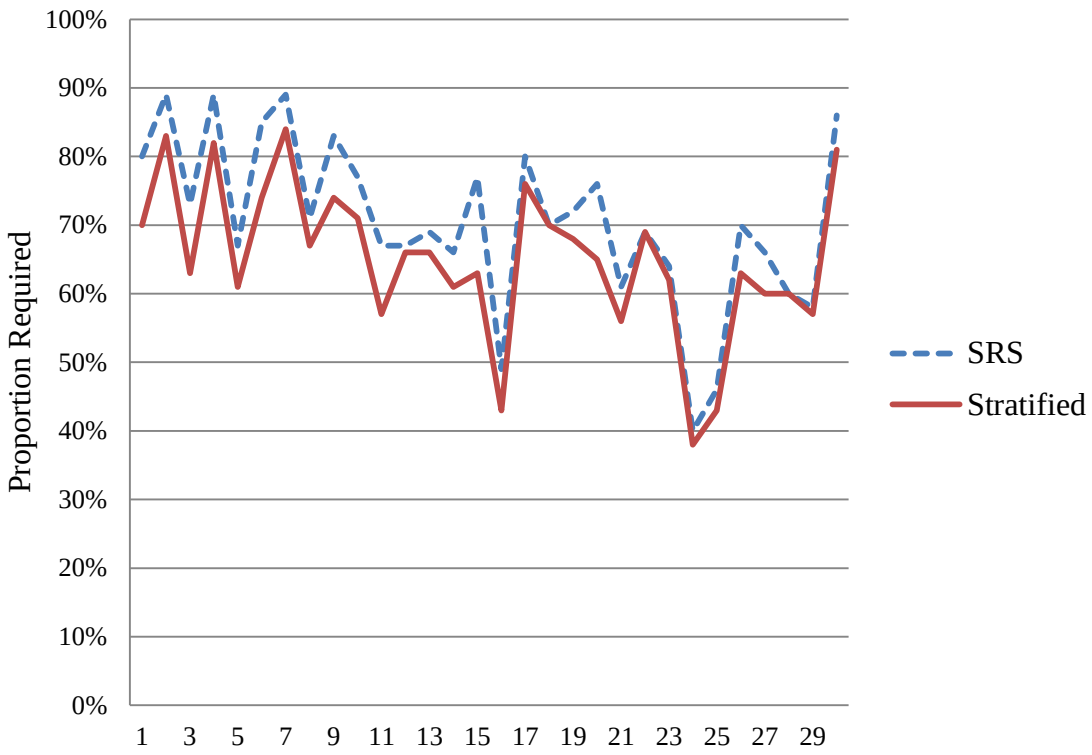

Figure 2. *Proportion of Test Required for Standard Within One % of the Full-Length Test.* SRS= simple random sampling. Stratified = stratified item sampling

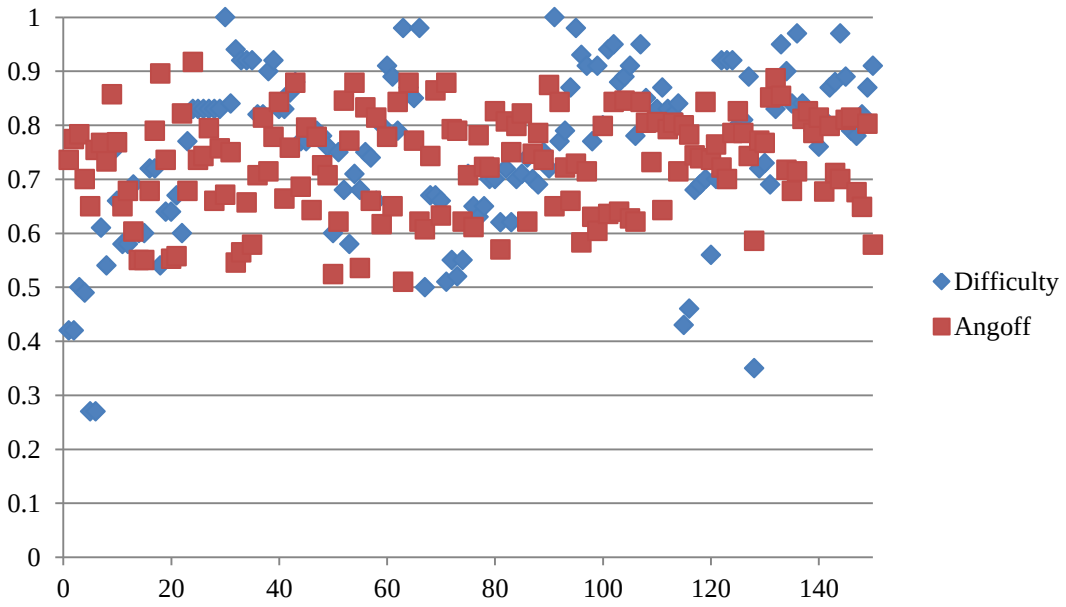

Figure 3. *Relationship Between Angoff Ratings and Item Difficulty Values for Items on Test 18.*

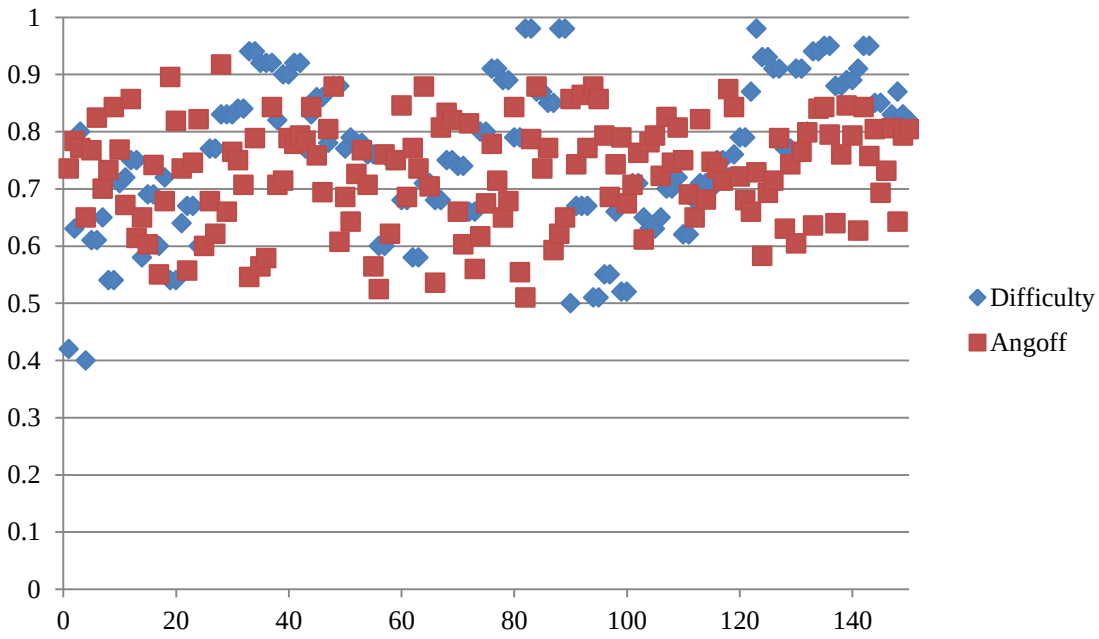

Figure 4. *Relationship Between Angoff Ratings and Item Difficulty Values for Items on Test 22.*

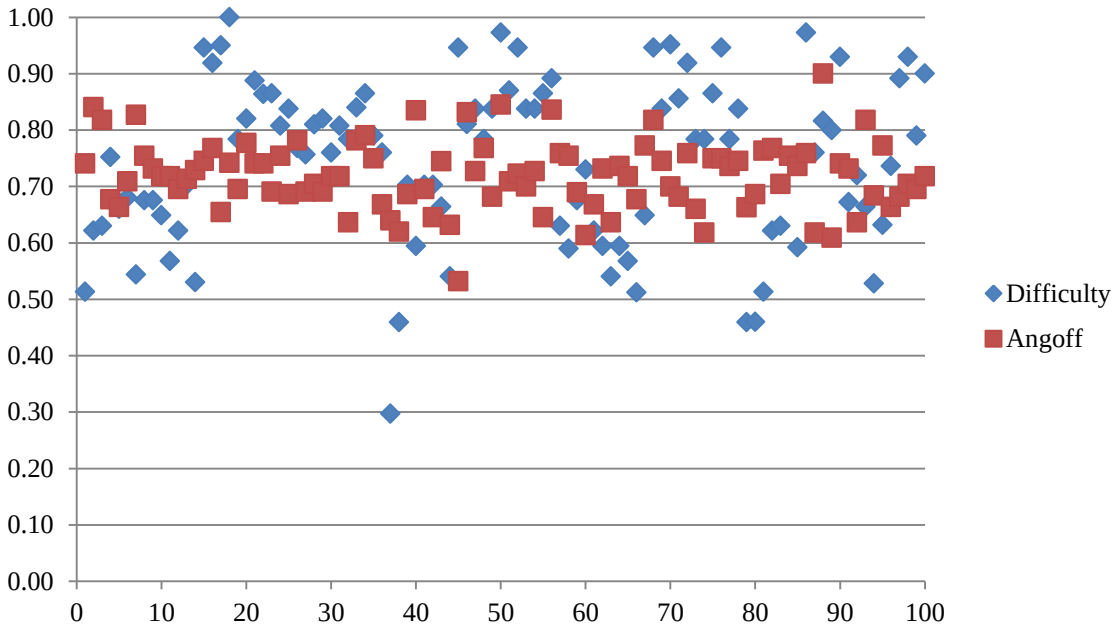

Figure 5. *Relationship Between Angoff Ratings and Item Difficulty Values for Items on Test 28.*

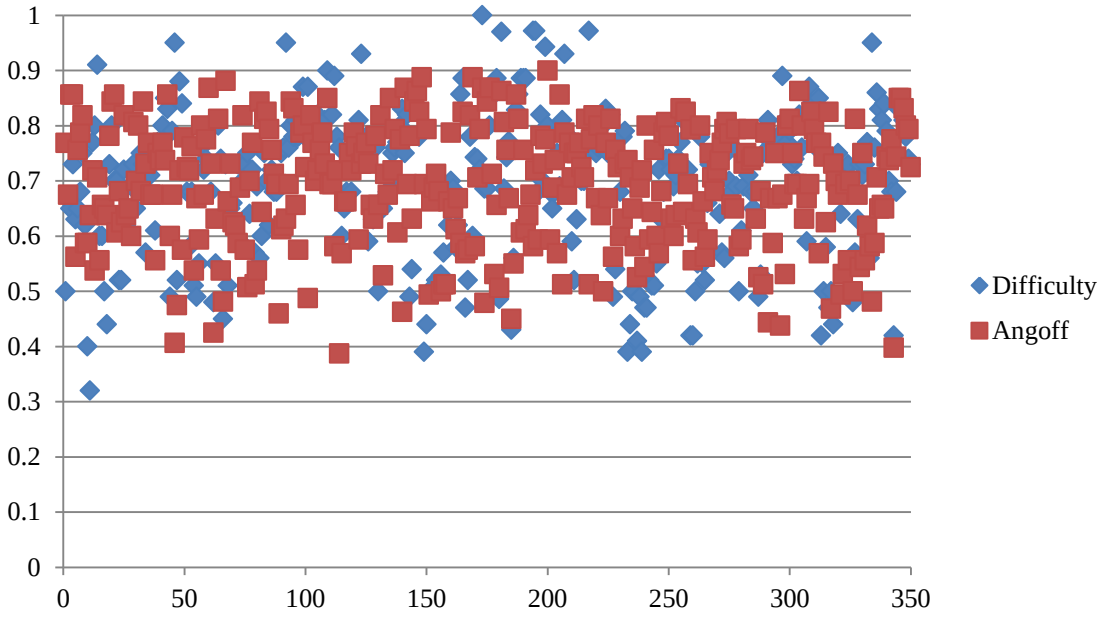

Figure 6. *Relationship Between Angoff Ratings and Item Difficulty Values for Items on Test 29.*

References

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

Behuniak, Jr., P., Archambault, F. X., & Gable R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. *Educational and Psychological Measurement, 42*, 247-255.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303-310.

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. (1996). Standard setting: The next generation of (where few psychometricians have gone before!). *Applied Measurement in Education, 9*, 215-235.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*, 59-88.

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). Judgemental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education, 12,* 151-165.

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

Cizek, G. J. (2006). Standard setting. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp 225-258). Mahwah, NJ: Erlbaum.

Downing, S. M. (2006). 12 steps for effective test development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of Test Development* (pp 2-25). Mahwah, NJ: Erlbaum.

Downing, S. M., Lieska, N. G., & Raible, M. D. (2003). Establishing passing standards for classroom achievement tests in medical education: A comparative study of four methods. *Academic Medicine, 78,* S85-S87.

Engelhard, Jr., G., & Cramer, S. E. (1992, April). *The influences of item characteristics on judge consistency within the context of standard-setting committees.* Paper presented at the annual meeting of American Educational Research Association, San Francisco, CA.

Fehrmann, M. L., Woehr, D. J., & Arthur, W., Jr. (1991). The Angoff cutscore method: The impact of frame-of-reference rater training. *Educational and Psychological Measurement, 51,* 857-872.

Ferdous, A. A., & Plake, B. S. (2005a). The use of subsets of test questions in an Angoff standard-setting method. *Educational and Psychological Measurement, 65,* 185-201.

Ferdous, A. A., & Plake, B. S. (2005b). Understanding the factors that influence decisions

of panelists in a standard-setting study. *Applied Measurement in Education, 18,* 257–267.

Ferdous, A. A., & Plake, B. S. (2007). Item selection strategy for reducing the number of items rated in an Angoff standard setting study. *Educational and Psychological Measurement, 67*(2), 193-206.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59,* 315-328.

Geisinger, K. F. (1991). Using standard setting data to establish cutoff scores. *Educational Measurement: Issues and Practice, 10,* 17-22.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-261.

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations.

Measurement: Interdisciplinary Research & Perspective, 2, 61–103.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences* (pp. 367-396). Berkeley, CA: McCutchan. Harker, J. K., & Cope, R. T.

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

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard setting. *Evaluation and the Health Professions, 6*, 3-24.

Harvey, A. L., & Way, W. D. (1999, April). *A comparison of web-based standard setting and monitored standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Hertz, N. R., & Chin, R. N. (2002, April). *The role of deliberation style in standard setting for licensing and certification examinations.* Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans: LA.

Horn, C., Ramos, M., Blumer, I., & Maduas, G. (2000). Cut scores: Results may vary. *National Board on Educational Testing and Public Policy Monographs*, *1*(1). Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

Hurtz, G. M., & Auerbach,M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*, 584–601.

Impara, J. C. (1995). *Licensure testing: Purposes, procedures, and practices*. Buros-Nebraska Series on Measurement and Testing: Buros Institute.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 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. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. *Applied Measurement in Education, 1,* 17-31.

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

Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice, 10,* 3-14.

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 status of validation in setting standards. In G. T. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (p55-88). Mahwah, NJ: Erlbaum.

Lamm, H., & Myers, D. G. (1978). Group-induced polarization of attitudes and behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11). New York, NY: Academic Press.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring.* Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998, April). *The Bookmark standard setting procedure: Methodology and recent implementations.*

Paper presented at the annual meeting of the National Council for Measurement in Education, San Diego, CA.

Linn, R. L. (1994, October). *The likely impact of performance standards as a function of uses: From rhetoric to sanctions.* Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Livingston, S. A. (1995). Standards for reporting the educational achievement of groups. *Proceedings of the joint committee on standard setting for large-scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES)*. Volume II. Washington, DC: National Assesssment Governing Board and the National Center for Educational Statistics.

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

McGinty, D. (2005). Illuminating the ‘‘black box’’ of standard-setting: An exploratory qualitative study. Applied Measurement in Education, 18, 269–287.

Mehrens, W. A. (1986). Measurement specialists: Motive to achieve or motive to avoid failure? *Educational Measurement: Issues and Practice, 5*, 5-10.

Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of the joint conference on standard setting for large scale assessments at the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II.* Washing, DC: U.S. Government Printing Office.

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education, 5,* 265-283.

Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement, 49,* 467-478.

Messick, S. (1995). Standards-based score interpretation: Establishing valid grounds for valid inferences. In L. Crocker & M. Zieky (Eds.), *In Proceedings of the joint conference on standard-setting for large-scale assessments,* Vol. 2 (pp. 291–305). Washington, DC: U.S. Government Printing Office.

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hasche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I* (pp. 73-85). Washington, DC: Council of Chief State School Officers.

Mills, C. N., & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Future of selected methods. *Applied Measurement in Education, 5,* 265-283.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement: Issues and Practice, 20,* 283-292.

Myers, G. D., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83,* 602-627.

National Assessment Governing Board. (1994). *Setting achievement levels on the 1994 National Assessment of Educational Progress in Geography and in US History and the 1996 National Assessment of Education Progress in Science.* Final version. Washington, DC: Author.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Norcini, J., Lipner, R., Langdon, L., & Strecker, C. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24,* 56-64.

Norcini, J., & Shea, J. (1992). The reproducibility of standards over groups and occasions. *Applied Measurement in Education, 5,* 63-72.

Norcini, J., Shea, J., & Kanya, D. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25,* 57-65.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., Koller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35,* 95-101.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education, 10*(1), 39-59.

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.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement, 15,* 297-230.

Raymond, M. R., & Reid, J. B. (2001).Who made thee a judge? Selecting and training participants for standard setting. In G. T. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (p55-88). Mahwah, NJ: Erlbaum.

Reckase, M. D. (2005, April). *A theoretical evaluation of an item rating method and a Bookmark method for setting standards.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10,* 11-14.

Scheaffer, R., Mendenhall, W., & Ott, L. (1979). *Elementary Survey Sampling.* North Scituate, MA: Duxbury Press.

Schulz, E. M., & Mitzel, H. C. (2005). *The Mapmark standard setting model.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Shephard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447-467.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36,* 477–481.

Sireci, S. G., Patelis, T., Rizavi, S., Dillingham, A. M., & Rodriguez, G. (2000, April). *Setting standards on a computerized-adaptive placement examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25*(4), 259-274.

Stoner, J. A. F. (1961). A comparison of individual and group decisions involving risk. Unpublished master's thesis, Massachusetts Institute of Technology, School of Industrial Management, Cambridge, MA.

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.

Wang, L., Pan, W., & Austin, J. T. (2003, April). *Standards-setting procedures in accountability research: Impacts of conceptual frameworks and mapping procedures on passing rates.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980's. In G. L. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19-51). Mahwah, NJ: Erlbaum.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). Cutscores: *A manual for setting standards of performance on educational and occupational tests.* Lexington, KY: Educational Testing Service.

.