THE IMPACT OF EXAMINEE PERFORMANCE INFORMATION ON JUDGES'
CUT SCORES IN MODIFIED-ANGOFF STANDARD SETTING EXERCISES

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

by
Melissa J. Margolis
August 2011

Examining Committee Members:

Joseph DuCette, Advisory Chair, Educational Psychology
Frank Farley, Educational Psychology
William Fullard, Educational Psychology
S. Kenneth Thurman, Educational Psychology
Brian Clauser, External Member, National Board of Medical Examiners

ABSTRACT

Background: Providing examinee performance data is a common procedural modification in modified-Angoff standard setting exercises. Results of research examining the impact of the practice are mixed, and despite the frequency with which this procedural modification is done, the overall impact on the outcomes is not well understood. This research presents a large-scale evaluation of the impact of providing performance data in the context of operational standard setting exercises for a medical licensing examination program.

Methods: Data from a total of 18 independent standard setting panels across three different examinations were analyzed to evaluate whether and how the provision of performance information impacted the resulting cut scores.

Results: Significant cut-score changes were found between initial and final judgments for all panels. A decrease in standard deviation between initial and final judgments for all panels was indicative of the convergence of judgments following the provision of data, and a lack of systematic changes in final cut scores suggested that judges were not simply changing their content-based judgments to align with the normative data. Both findings provide important procedural validity evidence for the standard setting process.

Implications: The current data set (replications of the procedure within a single cycle, across cycles for an individual examination, and across multiple examinations) allowed for the most comprehensive evaluation of this topic that has been described to date and provides valuable insight into an issue that is relevant to many testing contexts.

ACKNOWLEDGMENTS

Completing a Ph.D. is not something that one does in isolation, and I therefore would like to take this opportunity to express my gratitude to all of the people who supported me throughout this journey. Beginning with my family, the first heartfelt thank you goes to my parents Marilyn and Gene for their constant love, encouragement and support. Through the years they have never let an opportunity (or a greeting card!) go by without telling me how much they love and believe in me, and their words have fostered in me the knowledge that if I believe in myself and work hard I can accomplish anything. I also would like to acknowledge my sisters Gail and Jill and their families; I am blessed to have such a wonderful support system, and knowing that I have them behind me makes every endeavor just a little bit easier to endure. A huge thank you goes to my partner Lauren; without her love and support, this entire journey would have been immensely more difficult. Your unwavering patience, understanding, encouragement, and sacrifice allowed me to continue working to reach my goal even when it seemed like a near-impossible task. Thank you for believing in me and for helping me to continue believing in myself...your support made this accomplishment possible, and I am eternally grateful for everything you did along the way to help make this dream a reality. My thanks also go to Kathy and Chuck Entrekin... over the years you have been an extremely positive reminder of the fact that working, having a family, completing a doctorate, and remaining sane are not necessarily mutually exclusive endeavors! I also want to acknowledge my funny, loving, wonderful dogs as an important part of this experience: My dearest angel

Sasha, my crazy wayward Misha, and my sweet puppy Rex all dealt with different aspects of time, schedule, and attention constraints as I worked towards this goal, and along the way their crazy antics and unconditional love provided a [furry!] therapeutic diversion from the stress of my studies. And finally, to my son Bennett: I started on this journey long before you were born, but the journey ends with you as my inspiration. You are my heart and my soul.

I would be remiss if I didn't take a moment to express my gratitude to all of the friends who have shared in this journey and who now share in my success. I am especially grateful to Brian Clauser; you have been my supervisor, my mentor, and my dear friend, and I cannot imagine having gone through this process without your guidance, wisdom and unfailing support. Special thanks also go to Polina Harik who seemed nearly as dedicated to getting me to the end of this road as I was! Without your analytic wizardry, cognitive psychological insights, and supreme sign-making ability, I would not be where I am today. Thank you to Marci Winward, because coordinating the data that I used for my dissertation research would have been exponentially more difficult without your exemplary organizational skills and attention to detail! To the rest of my MCS family: I couldn't ask for a nicer and more supportive group of colleagues and friends, and the fact that you all were so genuinely invested in my success and were so excited [and nervous!] for me when I finally reached the day of my oral defense really meant the world to me. I'm glad that I didn't let you down! Finally, I will always be grateful that my graduate studies led me to cross paths with Christine Moran. You are a real-life Superwoman, and knowing you has been a constant reminder that I had no

option other than to finish what I started! Your support and encouragement all along the way have helped to keep me centered (I have three words for you...!), and I am extremely thankful to have you as an important part of my life.

I also would like to express my sincere appreciation to the National Board of Medical Examiners for its support of my doctoral work. After more than 17 years of employment, I still feel grateful to have found a professional home at such a noble and wonderful organization. Special thanks go to Ron Nungester for his patience and belief in me; I hope that you can find something else to fill the time you used to spend wondering when I would finish my degree!

Finally, I would not have reached this point without the support of my dissertation committee chair Dr. Joseph DuCette. Thank you for your dedication, encouragement and patience throughout this long process, and thank you for providing helpful feedback that immensely improved the quality of my dissertation. Thanks also go to Dr. Frank Farley, Dr. William Fullard, Dr. Kenneth Thurman, and Dr. Brian Clauser for agreeing to serve on my examining committee. Your participation was much appreciated, and your shared insights have helped me to think more critically about my research.

It is not possible to express my appreciation to every person who has been part of my life -and by extension a part of the process of getting my doctorate- and for that I apologize. Each person who touched my mind, my heart, and my life during this experience had a role in my success, and for that, I thank you.

TABLE OF CONTENTS

	Page
ABSTRACT.....	ii
ACKNOWLEDGMENTS	iii
LIST OF TABLES	viii
LIST OF FIGURES	x
CHAPTER	
1. INTRODUCTION	
Statement of Purpose	1
Rationale	1
General Methodology	6
Study Design	7
2. LITERATURE REVIEW	
Introduction.....	8
Standard Setting	9
Judgmental Standard Setting Procedures.....	12
Examinee-centered standard setting approaches	12
Test-centered standard setting approaches.....	14
Modifications to the Angoff Procedure	18
Validity and Standard Setting	25
Purpose of the Present Research	32
3. METHOD	
Introduction.....	34
The United States Medical Licensing Examination.....	34
USMLE Standard Setting Process	36
Content-based Standard Setting.....	37
Background	39
Process Training.....	40
Practice Judgments.....	41
Data	44
Analysis.....	44

TABLE OF CONTENTS

	Page
4. RESULTS	47
5. DISCUSSION	61
REFERENCES CITED	68
APPENDICES	
A. Sample Performance Information	74
B. Tables	76
C. Figures	89

LIST OF TABLES

Table	Page
4.1. Step 1 Panel-level Descriptive Statistics for Initial and Final Judgments	48
4.2. Step 2 Panel-level Descriptive Statistics for Initial and Final Judgments	49
4.3. Step 3 Panel-level Descriptive Statistics for Initial and Final Judgments	50
4.4. Mean Initial and Final Cut Scores and Associated Percents of Failing Examinees for All Examinations and Standard Setting Years	60
B1. Summary of Overall Dataset Structure	77
B2. Mean Cut Scores for Each Standard Setting Year Calculated by Taking the Mean of the Individual Panel Means (N=3)	78
B3a. Step 1 Year 1 Repeated Measures ANOVA Results: Tests of Within-subjects Effects	79
B3b. Step 1 Year 1 Repeated Measures ANOVA Results: Tests of Between-subjects Effects	79
B4a. Step 1 Year 2 Repeated Measures ANOVA Results: Tests of Within-subjects Effects	80
B4b. Step 1 Year 2 Repeated Measures ANOVA Results: Tests of Between-subjects Effects	80
B5a. Step 2 Year 1 Repeated Measures ANOVA Results: Tests of Within-subjects Effects	81
B5b. Step 2 Year 1 Repeated Measures ANOVA Results: Tests of Between-subjects Effects	81
B6a. Step 2 Year 2 Repeated Measures ANOVA Results: Tests of Within-subjects Effects	82
B6b. Step 2 Year 2 Repeated Measures ANOVA Results: Tests of Between-subjects Effects	82

LIST OF TABLES

Table	Page
B7a. Step 3 Year 1 Repeated Measures ANOVA Results: Tests of Within-subjects Effects	83
B7b. Step 3 Year 1 Repeated Measures ANOVA Results: Tests of Between-subjects Effects	83
B8a. Step 3 Year 2 Repeated Measures ANOVA Results: Tests of Within-subjects Effects	84
B8b. Step 3 Year 2 Repeated Measures ANOVA Results: Tests of Between-subjects Effects	84
B9. Step 1 Initial and Final Cut Scores and Associated Percent Failing by Panel.....	85
B10. Step 2 Initial and Final Cut Scores and Associated Percent Failing by Panel.....	86
B11. Step 3 Initial and Final Cut Scores and Associated Percent Failing by Panel.....	87
B12. Mean Initial and Final Cut Scores and Associated Percents of Failing Examinees for All Examinations and Standard Setting Years.....	88

LIST OF FIGURES

Figure	Page
4.1. Step 1 Year 1 Panel 1 Mean Cut Scores for Initial and Final Judgments.....	53
4.2. Step 2 Year 1 Panel 2 Mean Cut Scores for Initial and Final Judgments.....	54
4.3. Step 2 Year 2 Panel-level Mean Scores for Initial and Final Judgments	56
4.4. Step 3 Year 2 Panel-level Mean Scores for Initial and Final Judgments	57
C1. Step 1 Year 1 Panel 1 Individual Judge Initial and Final Cut Scores	90
C2. Step 1 Year 1 Panel 2 Individual Judge Initial and Final Cut Scores	91
C3. Step 1 Year 1 Panel 3 Individual Judge Initial and Final Cut Scores	92
C4. Step 1 Year 2 Panel 1 Individual Judge Initial and Final Cut Scores	93
C5. Step 1 Year 2 Panel 2 Individual Judge Initial and Final Cut Scores	94
C6. Step 1 Year 2 Panel 3 Individual Judge Initial and Final Cut Scores	95
C7. Step 2 Year 1 Panel 1 Individual Judge Initial and Final Cut Scores	96
C8. Step 2 Year 1 Panel 2 Individual Judge Initial and Final Cut Scores	97
C9. Step 2 Year 1 Panel 3 Individual Judge Initial and Final Cut Scores	98
C10. Step 2 Year 2 Panel 1 Individual Judge Initial and Final Cut Scores	99
C11. Step 2 Year 2 Panel 2 Individual Judge Initial and Final Cut Scores	100
C12. Step 2 Year 2 Panel 3 Individual Judge Initial and Final Cut Scores	101
C13. Step 3 Year 1 Panel 1 Individual Judge Initial and Final Cut Scores	102
C14. Step 3 Year 1 Panel 2 Individual Judge Initial and Final Cut Scores	103
C15. Step 3 Year 1 Panel 3 Individual Judge Initial and Final Cut Scores	104
C16. Step 3 Year 2 Panel 1 Individual Judge Initial and Final Cut Scores	105

LIST OF FIGURES

Figure	Page
C17. Step 3 Year 2 Panel 2 Individual Judge Initial and Final Cut Scores	106
C18. Step 3 Year 2 Panel 3 Individual Judge Initial and Final Cut Scores	107
C19. Step 1 Year 1 Panel-level Initial and Final Cut Scores.....	108
C20. Step 1 Year 2 Panel-level Initial and Final Cut Scores.....	109
C21. Step 2 Year 1 Panel-level Initial and Final Cut Scores.....	110
C22. Step 2 Year 2 Panel-level Initial and Final Cut Scores.....	111
C23. Step 3 Year 1 Panel-level Initial and Final Cut Scores.....	112
C24. Step 3 Year 2 Panel-level Initial and Final Cut Scores.....	113

CHAPTER 1: INTRODUCTION

Statement of Purpose

The purpose of the present study was to systematically investigate the practice of providing examinee performance data to content-based standard setting judges. Though this methodological approach has become relatively common in the context of standard setting for credentialing examinations, few rigorous investigations of the practice exist that support its use. This study highlights the importance of procedural evidence for standard setting practices -particularly for high-stakes examinations- and it sought to emphasize the contribution of rigorous decision-making practices to the overall validity of the examination.

Rationale

The process of setting standards and establishing appropriate cut scores is a critical step in any context that uses test scores to make decisions about examinees. Test-based decisions are made in a variety of areas such as high school graduation or equivalency, salary increases for teachers, course placement, and entry into a number of professions such as medicine, law and accounting. Though it may be widely agreed that setting standards is an important part of testing, how standards and cut scores are established often does not receive widespread attention from individuals other than those who are involved in the test construction process.

Establishing cut scores is not simply an important independent step in the testing process; this activity is critical with respect to the overall validity of the test (Kane, 1992, 1994, 2001, 2006). Validity has long been a topic of central importance in educational

and testing contexts. Although the importance of test (or score) validity has been recognized for generations, understanding of the concept has evolved over time. In the early part of the last century, the main use of scientifically developed tests was for job placement. As such, validity was a function of the ability of the test score to predict how well a person would perform a task or learn a job. In the middle of the last century, as tests were developed to assess characteristics for which there was no external criterion (e.g., introversion), construct validity became the popular model. This evolution in meaning has led to an understanding of validity not as a property of tests or test scores, but instead as a property of the interpretations or decisions that are to be made based on test scores. A recent conceptualization of validity highlights this distinction and focuses on the need to develop an interpretative argument which reflects the chain of evidence leading from test scores to score-based decisions. The extent to which an interpretation is valid is a function of the evidence that is provided in support of that specific interpretation. The more evidence that is collected in support of the interpretations that will be made about test scores, the more confidence a test user can have that those interpretations are appropriate (Kane, 1992, 1994, 2001, 2006).

According to Kane's conceptualization, there are four main components of the validation argument: Scoring, Generalization, Extrapolation and Decision. These components relate to separate yet overlapping aspects of the process, and evidence for each of them must be collected to ensure that the inferences that will be made based on test scores are, in fact, valid. Information relevant to the *Scoring* component would include evidence that the test was administered and scored properly, whether it is a

multiple choice examination that is administered and scored via computer or a performance-based examination that is scored by expert raters. The *Generalization* component is concerned with evidence that the scores are reliable (i.e., the test and testing conditions are such that an examinee taking a different but equivalent form of the same test would receive a similar score). The *Extrapolation* component of the argument requires evidence that the examinee's responses in the test situation are relevant to the real-world settings in which the inferences are to be made. Finally, the *Interpretation/Decision* component is concerned with evidence that the process used to establish cut scores and make decisions about examinees was reasonable.

While evidence for the first three components is fairly straightforward to collect via sound test design, test administration, and research protocols, collecting evidence for the last component may prove more challenging. This limitation may be particularly problematic in high-stakes testing contexts where the resulting pass/fail decisions can have enormous impact not only on the examinees, but, as is the case with credentialing (i.e., licensing or certifying) examinations, on the public as well. Receipt of a passing score on a medical licensing examination should provide an indication that the examinee has enough of the required knowledge and skills to be a safe and effective practitioner; should the procedures for establishing cut scores be called into question, the resulting decisions (e.g., the assumption that passing examinees are safe and effective practitioners) also must be questioned. Despite the clear importance of providing evidence for the validity of cut score decisions, relatively little research exists that systematically investigates the credibility of the processes that are used to produce these

decisions. As a testament to the importance of this topic, the 2006 National Council on Measurement in Education Career Award Address was dedicated to the subject of standard setting (Plake, 2008). In her address, Plake reviewed two decades of past research but clearly emphasized the need for additional research in this area.

The process used to establish cut scores will differ based on the specifics of the testing context and the personal preferences of the practitioner; there are many different procedures (and variations on those procedures) that are used in practice. Perhaps the most common approach to setting standards for multiple choice examinations is the Angoff method (Angoff, 1971). In the traditional Angoff approach, standard setting judges first are asked to conceptualize a “minimally proficient” examinee (the examinee whose level of proficiency justifies passing, but just barely). They then are asked to review test items and, for each one, to provide an estimate of the proportion of minimally proficient examinees that would answer the item correctly. These proportions then are aggregated across all of the test items and across all judges, and the resulting value provides an estimate of the score that a minimally proficient examinee would be expected to receive on the test (i.e., the cut score).

Over time, a number of modifications to the Angoff procedure have been suggested. One of the more common modifications involves providing judges with performance data for the items they review; this allows them to get a sense of how people taking the test actually perform on the items. The general purpose of the practice of providing performance data is to increase the reliability of the judgments and, by extension, to decrease the resulting error in the cut score (Hurtz & Auerbach, 2003). If

the result *is* a decrease in error, it would yield important evidence for the validity of the procedure that was used to establish the cut score. If, however, artifacts of using the data actually yield a cut score that is more a function of the provided data and less a function of the expert judgment that is expected to guide the process, the validity of the resulting cut score and subsequent decisions based on that cut score would be called into question.

A number of studies have investigated how the provision of performance data during the Angoff standard setting process impacts the resulting cut scores. Results of this research are mixed in terms of how the cut scores are affected; some studies found increases in cut scores, some found decreases in cut scores, and some found both increases and decreases across different content areas on the same test (Busch & Jaeger, 1990; Clauser, Swanson & Harik, 2002; Cross, Impara, Frary & Jaeger, 1984; Plake, Impara & Potenza, 1994; Truxillo, Donahue & Sulzer, 1996). The more uniform findings across studies were that providing data generally leads to convergence (i.e., group cut scores getting closer together; Busch & Jaeger, 1990; Clauser, Swanson & Harik, 2002; Cross, Impara, Frary & Jaeger, 1984; Plake, Impara & Potenza, 1994; Truxillo, Donahue & Sulzer, 1996).

One important limitation that interferes with straightforward interpretation of the findings of these studies is that the methodology often confounds the provision of performance data with other procedural modifications. One example of a common modification to the Angoff process is the inclusion of a group discussion component. This practice allows judges to hear their peers' Angoff judgments and to share the reasoning behind the judgments that they made. Though the general intent behind this

practice is a good one (i.e., to decrease the variability in the ratings), the problem from a research perspective is that this component of the process often is included with other manipulations and it therefore becomes difficult, if not impossible, to attribute any findings to a specific methodological approach. Results of these studies therefore must be reviewed carefully and the reported results must be interpreted with caution.

This dissertation expands upon past research and provides an important contribution to the literature by conducting a large-scale review of data that resulted from multiple high-stakes operational standard-setting exercises. The intent of the research was to provide evidence about the effects on cut scores of using a common modification on the Angoff procedure in which content experts are provided with examinee performance data. Though this modification has become common, the impact of providing such data is not fully understood and may have considerable implications for the validity of the resulting decisions. Results of this research provide important insight into the overall validity of the process that is used to establish cut scores, and, by extension, provide insight into the validity of interpretations that are made based upon the scores that result from the included examinations.

General Methodology

Data for this research were taken from operational content-based standard setting exercises that were conducted for the three examinations that comprise the United States Medical Licensing Examination® (USMLE). As part of its standard operational practices, a content-based standard setting exercise is conducted approximately every three years for each of the USMLE examinations. For each exercise, three independent panels of approximately 8-10 content experts are convened over a span of two-three months, and

each panel participates in an Angoff standard setting process using the same methodology.

Study design. Data were collected from two operational content-based standard setting meetings for each examination, and the data from each of the meetings were collected from three independent panels. During the Angoff process, judges reviewed items without performance information and provided judgments about the proportion of *minimally proficient examinees* that would get the item correct. Judges then were provided with two kinds of performance data: 1. Total test performance deciles that indicated the percentage of examinees at each decile that got the item correct; and 2. Information on the percentage of examinees that selected each multiple-choice option. They were asked to review the information, to determine whether something in the data suggested that they should make a change to their initial judgment, and then to provide a final judgment. The Angoff standard setting judgments that were made both before and after the judges saw the performance information and the cut scores that resulted from these judgments were analyzed for each examination. Descriptive statistics were calculated and repeated measures analyses of variance were conducted to investigate: 1. Whether the cut scores changed after judges received performance data; and 2. Whether any specific patterns of results had implications for the validity of the standard setting process. Additional analysis of the extent to which differences between pre-and post-data judgments impacted the resulting fail rates provided insight into the practical significance of the results.

CHAPTER 2: LITERATURE REVIEW

Introduction

The field of standard setting for educational assessments is extremely broad, and much literature describing relevant conceptual and practical issues has been published over the past 40 years. The present research focused on investigating the impact and implications of specific procedural modifications in one of the most popular judgmental standard setting methods: the Angoff procedure. To provide a context for this research, it first is necessary to provide a general introduction to the topic of judgmental standard setting procedures; such an introduction will allow for a broader understanding of these approaches and their respective strengths and weaknesses. The Angoff procedure then will be described in greater detail as the procedure that is the main focus of the present research, and literature on the specific procedural considerations that led to the present study will be elaborated and the specific research questions will be introduced.

While the main focus of this research was on standard setting and the decisions that result from judgmental procedures, validity issues provide an overarching context for all of this work. Providing validity evidence for a testing program involves consideration of a number of different aspects of the testing process above and beyond the specific test content. It also is important to provide evidence for the validity of decisions that are made as a result of the testing process. This includes collecting evidence to support the validity of the judgmental process that is selected and how that process is implemented. As such, a specific discussion of validity considerations in the context of judgmental standard

setting is included and will provide a framework within which to evaluate the outcomes of the research.

Standard Setting

The issue of standard setting is critical for contexts in which test scores are used to make decisions about examinees. In many educational and occupational testing settings, inferences about examinees are made based on how they perform on the test relative to other examinees; this type of test is considered to be norm-referenced (AERA, APA, NCME, 1999; Ford, 2010). In other contexts, the inferences of interest are based not on how examinees perform relative to one another but instead on how they perform in relation to previously defined standards for performance. In either context, specific points on the score scale -called cut scores- may be identified and used to distinguish between subgroups of examinees. If a test is used for selection purposes, for example, a cut score may be established that either passes or fails a pre-specified percentage of examinees; an interpretation that considers how examinees performed relative to other examinees in the testing group is a norm-referenced cut score interpretation. Alternatively, cut score(s) that are set to distinguish between qualitatively different levels of performance are criterion (rather than norm)-referenced; this determination is made based on the comparison of a performance to a pre-specified cut score (e.g., one that distinguishes between passing and failing performance), and in theory, all examinees could pass or all examinees could fail. The type of assessment and the type of interpretations that score users want to make based upon test scores will inform decisions about setting cut scores; in criterion-referenced contexts, where qualitatively different categories of performance are

identified, some approach to making informed and defensible decisions about where the cut scores should be must be employed; judgmental standard setting methods typically are used to inform this process. These methods necessitate that people with knowledge of both the test content and the population of examinees go through a structured exercise that allows them to provide insight into the level of performance that an examinee must achieve in order to demonstrate that the required level of proficiency has been reached. Though in some instances practitioners may treat this process as science, in truth it is truly what the name suggests: a way for experts to provide judgment and insight into the knowledge and skills that are required for inclusion in a particular performance category on a particular assessment. Though on the surface this might seem like a relatively straightforward concept, there are a few issues that warrant clarification at the outset of a long treatment of issues related to standard setting.

A discussion of standard setting in the fullest necessitates making a distinction between two steps in the process: setting the performance standard(s) and setting the cut score(s). Though these steps are often described in a manner which makes them seem overlapping or interchangeable, they are in fact distinct. The first step involves identifying the levels of performance that are being defined, whether there are two (e.g., unacceptable/acceptable, fail/pass) or more than two (e.g., unacceptable, basic, proficient, advanced). This step is critical regardless of the context; the question is one of how much a test taker must know or be able to do in order to be included in a given level of the competency of interest (e.g., knowledge of 3rd grade math, ability to safely fly and land an aircraft, ability to safely and effectively care for patients in supervised medical

practice). A straightforward way to think about the performance standard is that it is a conceptual representation of the desired level of competence (Kane, 1994).

The second step in the standard setting process involves collecting information that will inform setting the specific cut scores; these cut scores are the points on the actual score scale that are associated with the levels of performance defined by the performance standards. While the performance standards are conceptual representations of the desired level of competence, the passing score, or cut score, represents an operationalization of that minimally adequate level of performance for a given performance standard (Kane, 1994). The specific testing context will inform the specifics of how both of these steps are performed and how they relate to one another. For some tests, a single group of policymakers will make decisions about the performance standard and the passing score without external input. In other contexts, subject matter experts will act as standard setting judges and provide input into both the conceptual and the operational representations of required minimum levels of competence; policymakers then will make the ultimate decision about the point on the score scale that will represent the minimum passing score.

For the purposes of this dissertation, the specific focus was on the second step in the standard setting process, i.e., the one that involves collecting the data to inform the cut score decision. A variety of procedures are used to guide this process, and an overview of these procedures will provide important background and context for the specific procedure that is the focus of the present research. The following section presents such an overview of the most commonly used judgmental standard setting procedures.

Judgmental Standard Setting Procedures

Judgmental standard setting procedures historically have been divided into two categories: examinee-centered and test-centered approaches (Cizek, 1996a; Hambleton & Pitoniak, 2006; Kane, 1994, 1998). As the names suggest, examinee-centered approaches are those that focus on the examinees by asking people who know them (e.g., teachers) to place them into performance categories; this generally is done: 1. by considering how the examinees have performed in the past; and 2. without knowledge of performance on the test of interest (Hambleton & Pitoniak, 2006, Kane, 1994, 1998). In contrast, test-centered approaches involve the review of actual test material rather than review of the performance of a specific group of examinees; panelists review individual test items and make judgments about the expected level of performance for examinees who are between two levels (e.g., acceptable vs. not acceptable). The following section will provide a brief overview of the most common approaches associated with each of these two main categories.

Examinee-centered standard setting approaches. Examinee-centered approaches are used in both credentialing and educational contexts; in either case, they require the review of examinee performances and evaluation of the performances relative to the defined performance standards (Kane, 1998). With these approaches, judges are asked to make classification (e.g., pass/fail) decisions about test takers independent of the examination of interest; the location of the cut score then is determined by identifying the point on the score scale that is most consistent with those decisions. The borderline-group method (Livingston & Zieky, 1982) involves identifying individual examinees whose

level of achievement places them close to the performance standard; performance in this range is considered “borderline” because it is close to both the acceptable and the unacceptable ranges. Examinees may be categorized as borderline based either upon judges’ knowledge of their past performance or based on another metric; once this borderline group has been identified, their scores on the test of interest are calculated and the median score typically is used as the cut score. The main advantage of the borderline method is that it is conceptually a simple and straightforward exercise for judges; they understand and have little difficulty performing the required task. Despite that strength, there are several documented limitations of this method. The first is that categorization is dependent upon the knowledge that teachers or other raters have about the students; in contexts where additional information is not as readily available, the borderline judgments may prove more difficult to make. In addition, it is not clear that the categorization as borderline is always a reflection of judges’ knowledge about the students; in some instances, students may be classified as borderline because the judges are unsure about the students’ level of knowledge and/or skills (Hambleton & Pitoniak, 2006). Finally, the number of examinees in the borderline group is often quite small, and determining a cut score based on small numbers may be problematic not only conceptually but also with respect to estimating the test score distributions (Mills, 1995).

The contrasting-groups method is another popular examinee-centered method that also involves consideration of examinee characteristics or metrics other than scores on the test of interest (Livingston & Zieky, 1982). Using some independent criterion or criteria, examinees are categorized into two groups: those who are clearly below the

standard and those who are above it. For example, teachers or other evaluators may review examinees' work on prior assignments in order to determine whether this work provides an indication that they are or are not above the performance standard that indicates competence (vs. lack of competence) in a particular domain. Score distributions then are produced for each group, and a cut score is selected that best discriminates between the two groups. Though also conceptually straightforward, one limitation of this method lies in the potential difficulty in identifying the point that best distinguishes between the two groups. In addition, the generalizability of the sample to future administrations can also be questionable (Hambleton, Jaeger, Plake, & Mills, 2000).

While the two described examinee-centered approaches are those that are most commonly used, test-centered approaches overall tend to be the more popular judgmental approaches to setting standards, particularly in the context of credentialing examinations (Kane, 1994). In contrast to the examinee-centered approaches, test-centered approaches involve the review of test items or other test material in order to make the necessary decisions about required levels of performance. Several different methods provide alternative approaches to the standard setting task; three of the most commonly used methods are described below.

Test-centered standard setting approaches. In the Ebel method, judges categorize test items along two dimensions: difficulty and relevance to the decision that is to be made (Ebel, 1972). The product of the number of levels of difficulty and relevance (e.g., low, medium, high difficulty; essential, important, acceptable, questionable relevance) yields a total number of categories, and judges then decide on the proportion

of items in each category that a borderline examinee would answer correctly. The total number of items in each category is multiplied by the proportion of items in each category that would be answered correctly by a borderline candidate, and the products are summed to yield the cut score. Some features of this method have been identified as being potentially problematic in terms of both the process and the results.

The first aspect of the Ebel procedure that is a concern is the nature of the panelist task (i.e., considering difficulty and relevance separately); some researchers feel that it is too difficult for panelists because the two dimensions are very highly correlated (Shepard, 1984). A second consideration is panelist buy-in; because empirical item difficulty information is available for many if not most testing contexts, panelists may consider this step in the process unnecessary (i.e., it is possible to calculate the actual proportion of examinees that get an item correct based on past examination data, so asking the panelists to make that judgment may seem unnecessarily burdensome (Hambleton & Pitoniak, 2006). Finally, an additional disadvantage of this method that has been noted relates to the focus on relevance and to the potential discovery that material that is not considered essential or even important may be included on the test (Cizek, 1996a). This would call into question the process used to develop the test and would, at best, at least temporarily derail the course of the standard setting activity.

An approach that was specifically designed for use with a multiple choice item format is the Nedelsky (1954) procedure. With this approach, judges review each multiple choice question and decide on the number of options a minimally competent examinee would be able to eliminate as incorrect choices. After those incorrect options

are excluded, the minimum pass level (MPL) for an item is calculated as 1 over the number of remaining options. These item MPLs are then summed across items to determine the passing score on the test. As with the Ebel method, some features of the Nedelsky method have been identified as potentially problematic.

The first concern about the method has to do with the basic assumption underlying the procedure. This assumption asserts that the minimally competent examinee approaches answering test questions by eliminating clearly incorrect options and then randomly selecting a response from among the remaining options. There is, however, no clear evidence that random selection of a response is in fact how examinees ultimately approach the task, and the outcome of the process will be impacted by the extent to which the truth differs from this assumption (Jaeger, 1989; Shepard, 1984). A second limitation of this procedure is that the outcomes are directly linked to the number of multiple choice options offered in a test question. This not only provides a limit on the range of probabilities that can be produced, but it also tends to yield a preponderance of .50 probabilities (i.e., all but two options identified as clearly incorrect) which in turn leads to fairly low cut scores (Shepard, 1980, 1984).

The final test-centered approach that will be described is the Angoff method. It is generally considered the most widely used standard setting approach (Goodwin, 1999), and it is the procedure that is the focus of the present research. In its original form, the Angoff method asks judges to conceptualize a minimally competent examinee and to produce item-level MPLs by estimating the probability that the minimally competent examinee will answer each test item correctly. These probabilities are averaged across

judges to yield an overall item MPL and the item MPLs then are summed across items to yield the passing score. There are both strengths and weaknesses associated with use of the Angoff method. The major strengths are that it is a straightforward procedure that is easy to explain and to implement (Goodwin, 1999). In addition, the Angoff approach has been described as providing the best balance between the practical issues of implementation and the necessary technical rigor (Berk, 1986). Despite these notable strengths, there is an associated weakness in the procedure: judges' ability to conceptualize the minimally competent examinee. Judges are recruited to participate in the standard setting process because of their content expertise, their knowledge of the general or specific examinee population, or both; they are not recruited for their ability to estimate the probability of minimally competent examinees that would answer an item correctly. Because of this difficulty and the struggle that judges often go through as they're trying to understand: 1. how to conceptualize this minimally competent examinee; 2. how that conceptualization translates into test performance; and 3. what the specific associated probabilities are for a given set of items, efforts to provide judges with additional training and information to make their task more manageable have become common. These modifications to the originally described process have made some version of a *modified* Angoff procedure a more commonly used approach than the original. Though the basic logic of the Angoff method holds, a number of different procedural modifications have been applied to the original approach of simply estimating item-level MPLs. The majority of these modifications relate to the provision of performance data to the judges, and research to evaluate the performance of these

modified approaches tends to focus on the impact of this procedural modification on the resulting judgments and cut scores (Brandon, 2004; Busch & Jaeger, 1990; Clauser, Swanson and Harik, 2002; Cross, Impara, Frary, and Jaeger, 1984; Fehrmann, Woehr, and Arthur, 1991; Hurtz & Auerbach, 2003; Plake, Impara, and Potenza, 1994; Norcini, Shea, and Kanya, 1988; Truxillo, Donahue, and Sulzer, 1996).

Modifications to the Angoff procedure. Along with the widespread use of the Angoff procedure, there has been considerable interest in how modifications to the procedure, and in particular the presentation of different types of feedback, affect the resulting judgments and, by extension, the resulting cut scores. The importance of this body of research lies in the interpretations that can be placed on any identified procedure-based differences. As was mentioned earlier, experts are recruited to participate in content-based standard setting activities because of their knowledge of the content and their experience with the relevant population of examinees. If judges are abandoning their content-based judgments in lieu of the norm-based performance data they receive, the validity of the entire standard setting procedure may be called into question. Hambleton (2001) expressed the view that the impact of providing performance data “may be more psychological than psychometric” (p. 102) and asserted that the main impact often is on the variability among panelists rather than on the overall estimated cut score. In a subsequent review, Brandon (2004) reviewed six studies in which judges were provided with performance data and reported a significant change in the cut score in four of the six studies. Hurtz and Auerbach (2003) provided a meta-analysis of studies on the performance of the Angoff procedure and concluded that providing performance data

generally resulted in lowering the cut score. A number of additional studies also have addressed the impact of performance data on panel variability and on resulting cut scores.

In a 1990 study, Busch and Jaeger reported on standard setting exercises for seven subtests from the National Teachers Examinations. A single panel of between 25 and 50 members was used for each subtest, and the panelists made a initial judgment and then were given: 1. data indicating the proportion of examinees answering each item correctly; and 2. a cumulative frequency score distribution (for a defined examinee group). Results indicated that three of the seven cut scores changed by more than 0.5 points; two of these changes were statistically significant ($p < .05$). For the two subtests with significant cut score changes, one showed an increase and one a decrease. In each instance, the standard deviation (SD) of the cut scores decreased for the final judgment; this indicates reduced variability across judges and therefore that the panelists' judgments converged following the provision of performance data.

Plake, Impara, and Potenza (1994) examined whether content experts would differ from non-experts in the extent to which they used performance data to modify their judgments. English ($n=7$), science ($n=6$), math ($n=7$) and social studies ($n=6$) experts set standards for tests from the same four content areas. All experts met together for orientation and discussion of the minimally competent candidate; they then continued discussion in content-specific groups. After making an initial set of judgments, judges received three kinds of information: 1. proportion correct scores for each item; 2. proportion correct scores for examinees with total scores within 0.5 SDs of the temporarily established cut score; and 3. information reflecting the highest, lowest and

median rating of each item from within their content group. After receiving this information, judges had the opportunity to revise their judgments. Revised cut scores were higher than the initial cut scores for English and science items and lower than the initial cut scores for social studies items; for the math items, the cut scores were significantly different across groups and the direction of the change between initial and revised judgments varied across groups. As with the Busch and Jaeger (1990) study, the groups appear to have converged in their estimated cut scores after reviewing data. The group SDs also were lower for the final judgments than they were for the initial judgments; this indicates convergence within groups.

In another study that examined the impact of providing performance data on the cut score, Norcini, Shea, and Kanya (1988) asked a panel of six judges to review 48 items from a general certifying examination in internal medicine. Judges first reviewed and discussed the items. They then were given performance data (proportion correct scores for each item) and were permitted to change their judgments if they chose to. The results showed a (non-significant) 0.5 point reduction in the cut score between the initial and the revised judgments. The SDs decreased for all but one judge after receiving performance data, and across judges there was an overall decrease in SD following receipt of performance data that is suggestive of a convergence of judgments. One feature of this study that interferes with the ability to clearly interpret the results is that two different procedural modifications were combined: provision of performance data and discussion of the judgments among the judges. The initial judgments were collected verbally and posted on a blackboard, and judges had the opportunity to modify those judgments as part

of the initial discussion process. Following the provision of performance data, judges had another opportunity to discuss their judgments before arriving at a final decision. The final judgments may have been (and likely were) influenced by both the data and the group discussion, and therefore it is not possible to tease apart the effects of these two procedural modifications.

A study by Cross, Impara, Frary, and Jaeger (1984) presents the results of a complex study comparing several procedures, one of which was a variant on the Angoff procedure. Thirty faculty members were recruited to make judgments about the Mathematics and Elementary tests from the National Teacher Examinations; the overall group was divided in two so that 15 judges reviewed each of the tests. Judges completed a three-step process: 1. they reviewed the odd-numbered items without data; 2. they reviewed the even-numbered items with performance data reflecting the proportion of minimally competent examinees estimated to know the correct answer; and 3. they were asked to re-evaluate the odd-numbered items after reviewing two forms of performance data: the proportion of examinees estimated to know the correct answer and a cumulative frequency distribution which allowed them to see the impact of their initial estimated cut scores on a specific examinee population. For both the Mathematics and the Elementary tests, the mean cut scores dropped when judges had access to performance data (i.e., mean cut scores were lower for steps 2 and 3 than for step 1 of the review process). The SDs were lower for judgments made at step 2 than at step 1, which suggests that there was convergence following the provision of performance data, but for the Mathematics

test the SD was higher for step 3 than for step 1. That is, when judges re-evaluated the odd-numbered items with performance data, the agreement between judges declined.

Truxillo, Donahue, and Sulzer (1996) report on a sequential Angoff procedure applied to a test used as one hurdle in the selection of fire chiefs. Fourteen content experts participated in three panels that went through three iterations of an Angoff procedure (results are only reported by total group). Judges initially reviewed items and recorded their estimates of the probability of a correct response from a minimally proficient examinee. They then received performance data representing the proportion and number of examinees who answered the item correctly and had an opportunity to revise their judgments. Finally, they shared their judgments with the group and discussed each item; again they had a chance to revise their judgments. Across the three sets of ratings, the estimated cut score increased by more than 2.5 points (although this change was not statistically significant). The variability across judges also decreased after they received performance data and again decreased after discussion.

Clauser, Swanson and Harik (2002) examined results from a standard setting study for a medical licensing examination. Three independent panels of judges first provided initial judgments for a set of 75 items. They then were provided with two types of performance data for each item: “a histogram depicting the percentage of examinees answering the item correctly at each of 10 proficiency levels defined by decile on the total test score” (p.271) and the percentage of examinees that chose each of the options. They were asked to review the performance data, to use or not use the data as they chose to, and to provide a final judgment. For the initial judgments made in the absence of

performance data, there was considerable variability among estimated cut scores produced by the three independent groups of judges as well as those produced by judges in the same group. When performance information was provided: 1. judgments converged within and across groups; and 2. the estimated cut scores increased for two groups and decreased for the third.

Finally, Fehrmann, Woehr, and Arthur (1991) describe a study that reflects on the issue at hand without providing direct comparative results. An Angoff standard setting exercise was done for a 30-item general psychology multiple-choice test that was administered to 155 undergraduate students. Twenty-one psychology graduate students were recruited to act as expert judges and were randomly assigned to one of three panels. Individuals in the first group independently read instructions describing the required judgment and defining minimal competence (defined by the researchers as a student who would get a “C” in the course); they then provided judgments for the 30 items. Judges in the second group also acted independently. They were given the same instructions as those in the first group, but they were additionally provided with three sample items (not included in the set of 30) and were told that minimally competent examinees had a 90% chance of correctly answering the first item, a 50% chance of correctly answering the second item, and a 20% chance of correctly answering the third item; this information was intended to provide a frame of reference for making the item-level decisions. They then provided judgments for the same 30 items seen by the first group. The third group met together and discussed their definition of minimal competence (agreeing that it was associated with a “C” in the course). They provided judgments for the three sample items

and then were provided with normative data for those items. Following additional discussion, they independently rated the 30-item set. Results indicated that the two groups that were exposed to sample items had lower cut scores than the group that was not. These two groups also had significantly smaller total within group variances ($p < .05$). The variance components for items and for the residual were larger for groups two and three and the rater variance component was smaller. This suggests that convergence occurs because of the provision or development of a common frame of reference.

The above studies investigated and provide some general insight into the impact of providing performance information on the cut scores that result from Angoff standard setting exercises. They do not, however, provide much insight into the overall implications of the results either in terms of the specific procedures that were used or the overall meaning of the findings. The process of collecting information about the impact of this specific procedural modification is a necessary but not sufficient condition for providing evidence for the validity of both the standard setting and the overall examination process; knowing what the outcomes are is important, but addressing the implications of those outcomes is an important next step. The following section describes the role of validity in standard setting and addresses the importance of considering the implications of the outcomes of the standard setting process on the validity of the decisions that are being made based on the resulting cut scores and, by extension, on the overall validity of the examination.

Validity and Standard Setting

When considering test validity, it is common to think of factors related to methods of developing, administering and scoring tests. One step that often is overlooked in the overall consideration of validity is the importance of providing evidence for the ways in which the resulting scores are interpreted. This section provides some insight into this issue by starting with a brief historical context for the idea of test validity. The section then continues with a discussion of a more recent conceptualization of validity that makes clear the importance of collecting evidence for all aspects of the testing process including how cut scores are set and the appropriateness of inferences that are made about examinee performance based on the resulting testing decision.

Over time, the conceptualization of validity has undergone a dramatic evolutionary process. During the First and Second World Wars, the challenge of placing recruits was critical for the U.S. military and testing became a way of identifying appropriate job placements. Following the war, testing focused on improving industrial efficiency; in both cases, the issue of validity was centered around the prediction of job performance. As such, *criterion* validity became the common conceptualization of validity; the focus of this type of validity was on simple correlations between test scores and a criterion of interest. In this context, calculating the correlation between the test score and job performance was viewed as essential and quantifying the criterion was often reasonably straightforward. For example, candidates for pilot training school could be tested and the resulting scores could then be related to some aspect of their job performance after training. A test that could be shown to accurately predict which

candidates would successfully complete training and become pilots (or better yet, which candidates would perform successfully in combat) could be considered valid (Flanagan, 1948).

Over time, the science of testing expanded to support both achievement testing and the growing field of clinical psychology; this shift in focus produced measures of characteristics for which simple external criteria were not available (e.g., knowledge of American history, introversion). As a result, test developers were forced to reconsider the question of how their measures should be validated. In the area of achievement testing, evidence was gathered to demonstrate that the content of the test was consistent with the domain of interest. Such content validity became an important part of the validation process (although subsequent theorists have pointed out that this type of validity evidence provides no direct evidence about the interpretability of scores; Messick, 1989). In contrast to the achievement testing situation, the advent of psychological testing led to a need for assessment of constructs that were developed within a theoretical framework. In these situations there was no defined content domain from which test material could be sampled; it therefore was necessary to collect evidence indicating that the scores from these assessments were consistent with the associated theory. In response to this, Cronbach and Meehl (1955) proposed the concept of construct validity. This approach required a broader view of test or score validation and an expansion in the kinds of evidence that could be used for evaluative purposes. For example, in achievement testing the process of construct validation might include evidence that performance on the test of interest improves as the level of training increases. This shift in thinking has led to a

wider understanding of validity, and in 1989, Messick's chapter on validity in the third edition of *Educational Measurement* made clear what the trends were suggesting; instead of validity being considered a property of tests or test scores, it actually should be considered a property of the interpretations or decisions that are to be made based on test scores.

Kane (2006), in the subsequent edition of *Educational Measurement*, describes a conceptualization of validity that highlights the distinction initially described by Messick (1989). Kane's approach focuses on the entire testing continuum, from test development to the interpretations that score users are going to make about examinees based on their test scores, and highlights the need to collect evidence at many different points along the way. The purpose of focusing on the entire testing continuum is to develop what Kane calls an "interpretative argument" that reflects the chain of evidence leading from test scores to score-based decisions. Successful collection of a single piece of validity evidence is not sufficient to declare a test-based interpretation valid; because the validity of an interpretation is a function of the evidence that is provided in support of that specific interpretation, increasing amounts of evidence supporting the intended interpretations will lead to increased confidence that those interpretations are appropriate (Kane, 1992, 1994, 2001, 2006).

Four main components of the interpretative argument comprise Kane's conceptualization of validity: Scoring, Generalization, Extrapolation and Decision. Each component refers to a specific aspect of the testing process, and evidence for each must be collected to provide support for the validity of the inferences that will be made based

on test scores. While these components are described as being distinct, it is important to note that there are no rigid boundaries between them; one type of evidence may in fact relate to multiple components of the argument. Relevant to the *Scoring* component would be evidence that the test was administered properly and that it was scored properly. If the test was meant to be administered under standardized conditions, did those standardized conditions apply? Were the appropriate security procedures followed? Were the scores recorded accurately? Were scoring algorithms applied accurately? If the test was a multiple-choice examination, was the scoring key accurate and applied appropriately? If the test was a performance-based examination scored by expert raters, did the raters use the rating scales appropriately? All of these issues have the potential to impact the validity of score interpretations; if, for example, an examinee cheats on a test by gaining access to test material prior to the test administration and receives a passing score, it would be difficult to argue that a reasonable interpretation of the score in that instance is that the examinee has the necessary knowledge to meet the standard. Instead, advance access to test material should seriously call into question the validity of any interpretations that would be made about the resulting score.

The *Generalization* component is concerned with the extent to which replications of the measurement procedure would yield similar scores or classification (i.e., pass/fail) decisions. The validity of the scores would be called into question if there was a high likelihood of examinees receiving a completely different score or classification if they were to retake the examination. Relevant to the generalization component also would be evidence that the test had been constructed appropriately with respect to content and

sampling. The two main considerations with respect to this last issue are that the test content was selected in such a way that it is representative of the domain to which the test score is to be generalized, and that the overall sample of test material is sufficient. If the test is intended to assess the examinee's knowledge of the Civil War yet the majority of test items ask about the Battle of Bunker Hill, an interpretation of the score as a reflection of the examinee's Civil War knowledge would be called into question.

The *Extrapolation* component of the argument is concerned with evidence that the examinee's responses in the test situation are relevant to the real-world settings in which the inferences are to be made. Tests provide a way of collecting indirect evidence of an examinee's knowledge or ability in a specific area, and the main focus of the extrapolation component is on how indirect that evidence is. Evidence of a relationship between outcomes from the assessment of interest and other relevant measures can provide some evidence for this component. Similarly, evidence that the content of the examination is representative of the domain of interest also will provide some insight into the relevance of test performance to real-world performance. Evidence of the *absence* of sources of variability that are unrelated to the proficiency of interest provides another important contribution to the interpretative argument; these sources of variability are collectively referred to as construct-irrelevant variance. The assessment format can be a potential source of construct-irrelevant variance; with tests that are delivered via computer, familiarity with the computer interface, typing ability and timing constraints may exert an unintended influence on performance. Even in testing scenarios that are designed to closely approximate real-world settings, such as patient-care simulations in

which examinees interact with simulated patients, artificialities of the testing situation such as timing constraints and the way that the constructs of interest are scored can influence performance in ways that may call the score interpretation into question.

The final component of Kane's interpretive argument is *Decision*, and this component is concerned with evidence that the process used to establish cut scores and make decisions about examinees was reasonable. Evidence for the first three components is fairly straightforward to collect via sound test design, test administration, and research protocols, but the process of collecting evidence for the last component may prove more challenging. One of the main considerations within this component is the ability to provide evidence that the process(es) used to establish standards, to determine examination cut scores, and to make classification decisions were reasonable: this is referred to as procedural evidence for validity (Kane, 1994). When collecting this type of validity evidence, the areas of interest are the selection of the procedure, implementation of the procedure, and evaluation of the procedural evidence (Kane, 1994). There are no absolutes in judgmental standard setting, and even decisions relating to which procedure to use and the specifics of how it will be implemented are subject to the judgment of the decision makers. A careful and thoughtful approach to process selection and implementation and a subsequent evaluation of the outcomes can provide evidence for the interpretations that will be placed on the scores that result from the testing process. Alternatively, the identification of serious oversights or identified errors in the procedures can essentially invalidate an intended interpretation (Kane, 1994). If the provision of performance data as part of Angoff standard setting methods is intended to enhance the

process by increasing the reliability of content-based judgments but instead it serves to substantively change the nature of the judgments from content-based to norm-referenced, this would seriously call into question the procedures that are used to establish examination cut scores and by extension would call into question any interpretations that are made about examinees based on their test scores. As mentioned in the *Standards for Educational and Psychological Testing* (the *Standards*), “In some situations, the validity of score interpretations may hinge on the cut score” (AERA, APA, NCME, 1999, p. 53).

A serious flaw in the procedures used to set examination cut scores would be a concern in any testing context, but in high-stakes contexts where the resulting pass/fail decisions can have enormous impact not only on the examinees but also on the public, attention to this matter is even more critical. Receipt of a passing score on a medical licensing examination should provide an indication that the examinee has enough of the required knowledge and skills to be a safe and effective practitioner; if the cut score information resulting from judgmental standard setting procedures were called into question, the resulting decisions and interpretations (e.g., the assumption that passing examinees are safe and effective practitioners) also would be in question. Despite the clear importance of providing evidence for the validity of cut score decisions, relatively little research exists which is dedicated to systematically investigating the credibility of the processes that are used to produce these decisions. As a testament to the importance of this topic, the 2006 National Council on Measurement in Education Career Award Address was dedicated to the subject of standard setting (Plake, 2008). In her address,

Plake reviewed two decades of past research and clearly emphasized the need for additional research in this area.

Purpose of the Present Research

The present research was an attempt to address the identified need for contributions to the standard setting literature by investigating the impact of a common modification to the Angoff method – the provision of performance data – on the outcomes of multiple replications of a modified Angoff standard setting procedure. Results provide insight into common standard practices and the implications of those practices for the validity of decisions that are made based on the resulting cut scores.

The body of research described above has examined how judgments and cut scores change when content experts are given access to performance data. Though the researchers are to be commended for their independent investigations into this topic, the interested reader is still left wondering about the overall meaning of the body of reported results. The reports of increases or decreases in cut scores following the provision of data and the reports of convergence between and/or among independent groups provides some insight into the topic, but as an overall approach to interpreting the findings it is appealing to believe that the identification of some specific pattern(s) in the results would provide more conclusive support for (or evidence that undermines) the validity argument for cut scores produced with the Angoff procedure. Consideration of the previous research led to the identification of two main patterns of results that can provide insight into the implications for validity:

1. Convergence enhances credibility. Convergence of judgments by independent raters or groups may be viewed as useful evidence of procedural validity. At a minimum, it provides evidence that changes were not arbitrary or random and suggests that some “signal” was present that was interpreted similarly across raters.
2. Systematic shifts in judgments may argue against validity. A consistent increase (or decrease) of judges’ cut scores after review of performance data would imply that judges were using the content and performance information to make different decisions. This finding may suggest that the content-based decisions are systematically inconsistent with their norm-referenced expectations and further would suggest that the use of the data doesn’t reduce random errors in judgment but instead [inappropriately] realigns the intended content-based judgments with a systematically different norm-based judgment.

The present research used the above framework to evaluate the outcomes of 18 independent Angoff standard setting panels that participated in content-based standard setting exercises for the three distinct examinations that comprise the United States Medical Licensing Examination. This represents the most comprehensive data set that has been described and evaluated in the standard setting literature. The methodology is described in the following section and includes a description of the examination system that provided the specific context for the standard setting activities, a description of the specific procedures that were used during the actual content-based standard exercises, and a description of the analyses.

CHAPTER 3: METHOD

Introduction

Prior to describing the specific methodology associated with the present research it first is important to describe the background and context for the work that was done. This section includes background information on the examination system that provided the context for this standard setting research: the United States Medical Licensing Examination. The general standard setting approach used by the United States Medical Licensing Examination then is described and is followed by a description of the specific procedures that were used for the content-based panel meetings that provided the data for this study. The final section includes a description of the resulting data and the analyses.

The United States Medical Licensing Examination

The United States Medical Licensing Examination (USMLE) provides a single examination pathway for allopathic (i.e., MD) physicians in the United States. Licenses to practice are regulated and granted by the individual states and territories; anyone wanting to practice medicine in the United States must take and pass USMLE in order to be eligible to apply for a license. The USMLE process is divided into three independent Steps, each of which necessitates taking and passing one or more examinations. USMLE Step 1 comprises a single-day multiple-choice examination that assesses knowledge of the sciences that are basic to the practice of medicine; Step 1 generally is taken by students in United States (U.S.) medical schools at the end of the second year of medical school. Step 2 is comprised of two independent examinations: Step 2 Clinical Knowledge (CK) is a single-day multiple-choice examination that assesses the clinical knowledge

that is essential for the safe and effective practice of medicine under supervision. Step 2 Clinical Skills (CS) is a single-day performance-based examination that assesses the clinical skills that are necessary for safe and effective practice. Step 2 generally is taken by U.S. medical students in the fourth year of medical school. Though there are traditional times when U.S. medical students take the Step 1 and 2 examinations, they can be taken at any time and in any order; the only specific requirement for USMLE timing is that examinees must take and pass the Step 1 and 2 examinations before they can take the Step 3 examination. Step 3 is a two-day examination comprised of multiple-choice questions and computer-based case management scenarios. It generally is taken in the second year of post-graduate training and assesses application of medical knowledge and understanding of clinical science that are considered essential for the unsupervised practice of medicine. USMLE is jointly sponsored by the National Board of Medical Examiners (NBME) and the Federation of State Medical Boards (FSMB); members of both organizations comprise the Composite Committee, which is the decision-making group that has ultimate oversight of the entire USMLE process. Each of the Steps has its own individual oversight committee that is responsible for making examination-specific decisions.

The focus of the present research is on the process used to collect content-based judgments for the multiple-choice components of the USMLE examinations. For that reason, this work does not include the Step 2 CS examination or the computer-based case management component of the Step 3 examination.

USMLE Standard Setting Process

When USMLE was created, guidelines for how examination cut scores would be established and reviewed were specified. These guidelines indicated that a full review of standards should be conducted approximately every three years for each Step and that the full review would include the provision of several different kinds of data to the oversight committee. The cornerstone of the standards review process is to convene panels of subject matter experts to review and make judgments about actual test content. Results from this content-based standard setting process are the main focus of the full standards review and also are the focus of this dissertation. In addition to results from the content-based panels, the additional sources of data presented to the oversight committees are: 1. results of a large-scale survey of numerous constituency groups; and 2. examination-specific analyses that show: a. trends in examinee performance over time; and b. the impact on pass/fail rates of selecting different cut scores. These different sources of information allow the decision-making group to consider the issue of examinee performance from a variety of perspectives and to determine whether all of the sources of information together indicate that there is reason for making a change to the existing cut score (minimum passing score). The existence of this additional step in the cut score decision-making process is important to note. In much of the reported research, the cut scores that result from the described standard setting processes are those that are implemented operationally. In the examination system that is the focus of the present research (and, in fact, in most credentialing contexts) there is an additional step in the decision-making process. The content-based standard setting activities yield cut score

data, but those cut scores are not directly implemented as operational cut scores. Instead, the oversight group considers the resulting cut score information in addition to the other relevant sources of examination data and makes a determination about whether the cut score should change (and, if so, by how much). The data resulting from the content-based panels are in fact used to inform cut score decisions, but it is important to note the difference between this approach to establishing examination cut scores and one where the results of the content-based panels directly influence the new cut score.

Relevant to the main purpose of the dissertation, the focus of the remaining methodological description is on the content-based standard setting component of the overall standard setting process.

Content-based Standard Setting

For each of the USMLE Step examinations, the first step in the content-based standard setting process is to recruit qualified groups of experts to serve as standard setting panelists. Participants are recruited based on nominations from colleagues, and in addition to the requirement for teaching/supervising medical students/graduates, the panels are constructed according to several additional considerations. The first consideration is examination content. For example, because the content of Step 1 is rooted in basic science, Step 1 standard setting panels are comprised both of physicians and non-physician basic scientists (e.g., Ph.D. pharmacologists). The Step 2 and Step 3 committees are comprised entirely of physicians, and specific job responsibilities are further considered when recruiting for these panels. Because Step 2 assesses the readiness of examinees to practice in a supervised setting, attention is paid to ensuring that there is

a reasonable representation of individuals who have experience with examinees at the end of undergraduate medical education (when the test is taken) and those who have experience with examinees who are presently in the setting of interest (i.e., in post-graduate (residency) training). Additional recruitment considerations focus on creating panels that are diverse in terms of geographic representation, gender, ethnicity and medical specialty. Careful attention to and documentation of panel composition is one of the considerations mentioned in the *Standards*, and USMLE goes to great lengths to select participants and to create content-based panels that are both diverse within a given panel and similar in composition across panels (AERA, APA, NCME, 1999).

Due to the requirement to revisit the standards approximately every three years, one of the Steps convenes content-based panels each year; the procedures for implementation are the same for all three Steps. Because this process has been in place for a number of years and because the panel meetings occur on a regular basis, multiple datasets were available for analysis. This research analyzed data resulting from two different years of panel meetings for each of the three included Step Examinations. The Step 1 data are from a total of six independent panels, three of which were convened in 2003 and three of which were convened in 2006. The Step 2 CK data resulted from six independent panels, three of which were convened in 2002 and three in 2007, and the Step 3 data resulted from six independent panels, three of which were convened in 2004 and three in 2008. The following section describes the process that was used for all of the panels. (Note: In the standard setting literature, the individuals who are recruited to participate in content-based standard setting activities are referred to by a number of

different names. The terms that will be used interchangeably throughout this dissertation are panelist and judge; there is no distinction between them.)

Background. For each Step, three independent panels were assembled according to the recruitment criteria described above. Each panel was convened at the NBME offices in Philadelphia; panelists arrived for a 1:00pm meeting start and the meetings ended at approximately 3:00pm on the following day. One of the most critical pieces of the entire content-based standard setting processes is the training of the panelists; these individuals are recruited based on their professional expertise and their experience with the examinee population of interest, but there is no expectation that they will have had any experience with the specific standard-setting activity in which they were recruited to participate (AERA, APA, NCME, 1999; Hambleton & Pitoniak, 2006). As a result, the first full afternoon and the following morning of activities are viewed as training. To begin, panelists were provided with an overview of standard setting: what standard setting means, what the specific process was that they would be using, and how their judgments would translate into a cut score. They were told that *standard setting* is a matter of determining “how much is enough”, and when that question is applied to a testing context, the issue really is one of determining “to what level of performance should the lowest passing score be assigned?”. Beginning with this information provided a framework for the panelists and allowed them to understand in a general way what their role was and what the overall question of interest was for the entire process. The presentation then continued with some specific information about the Angoff standard setting process and the steps that they would be taking to make the judgments that were

being asked of them (that is, judgments about how the minimally proficient examinee would answer a given test question). The final piece of the introduction to standard setting informed the judges of how the individual item-level judgments that they would make would be translated into an actual cut score on the examination.

Following the standard setting introduction, judges were provided with some basic background on USMLE, including the purpose and structure of the examination sequence, and they then were provided with specific information about the test that they were there to review. This specific information included a detailed description of the purpose statement for the test and a broad overview of the content categories. The purpose statements provide information on the level of performance that is relevant to the decisions that judges are asked to make; focusing on that information provides judges with an initial framework for how to start thinking about their task. Similarly, providing some initial introduction to the content of the examination provided judges with a loose framework for what to expect when they began reviewing test items.

Process Training. Training on the actual content-based standard setting process began with a discussion of the concept of the *Minimally Proficient Examinee* (MPE); this is the examinee whose performance is *just acceptable* when measured against the standard of interest. (for example, on the Step 3 examination the MPE would be an examinee whose performance is acceptable enough to allow him/her to enter unsupervised medical practice). This is a critical step in the process, because understanding of and comfort with this concept is the foundation for all of the work that the panelists do. Panelists first were asked to think about the idea of the MPE and then to

try to describe the characteristics of this person. This is a notoriously difficult task, and guidance often was necessary to encourage the judges to begin to discuss their initial thoughts. The three main questions that panelists were asked to consider were: 1. Is there a specific kind of person who comes to mind?; 2. What are the characteristics of this person?; and 3. How will these characteristics impact test performance? Following initial discussion of the topic, some additional process-related details were provided and the training moved into the next phase: practice judgments.

Practice Judgments. For this initial exercise, judges were given a booklet containing 15 practice items and the associated answer key for those items. These practice items were selected to represent a range of item presentations that the judges would see throughout the main item set (i.e., the items were in a variety of formats including, but not limited to, differing numbers of options, associated images, associated tables or charts, etc.). The items also were selected to represent a range of difficulty levels so that the judges would have initial experience with providing estimates for and being able to discuss items of easy, moderate and high difficulty. The judges reviewed each item one at a time, provided a written judgment about their performance expectations, reported their judgments to the group and had them recorded on a whiteboard, and then discussed the reasons for providing the judgments they did. The question about performance expectations was asked in the following way: “What is the probability that a MPE would answer this item correctly?” (to facilitate understanding of the task, judges also were told that they could think of the question as “Out of a group of 100 MPEs, how many would answer this item correctly?”). Responses were recorded as

whole numbers, and judges were permitted to use any numbers they deemed appropriate (from 0 to 100).

After going through each item, providing initial judgments, and discussing the judgments with the group, judges were provided with data that indicated how a cohort of examinees had performed on each of the test items. Two distinct types of performance data were presented: 1. item-level graphs showing how examinees at each performance decile (based on total test score) performed on the item; and 2. a breakdown of the percentage of examinees that had chosen each of the multiple-choice options (see Appendix 1). Both pieces of information provided judges with an indication of how difficult the items were; the decile data indicate how difficult the individual items are in relation to how examinees perform on the overall test and the option data provide an indication of how attractive each of the incorrect options is to examinees. Judges were asked to review the performance data, review the item and their initial judgment again, and decide whether or not they thought that they should change their original judgment based on the data. The reason for providing the data is that research indicates that judges have difficulty making the required judgments without some sense of how examinees actually perform on the items (Busch & Jaeger, 1990; Clauser, et al., 2009; Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Clauser, Swanson, & Harik, 2002). Judges were instructed to use the data merely as a guide; if their judgment and the data were quite different, going back and taking a more careful look at the item could reveal that the judge missed something upon the initial review. The data were not intended to provide strong indications that the judges should change their judgments to match the data; they

were merely intended to provide a frame of reference. After reviewing the data for an item, judges were asked to provide a final judgment (which may or may not have been the same as the initial judgment) and again the judgments were recorded and discussed.

There are several reasons for including this practice judgment component of the training in the overall process. The first is that at this stage of the process the concept of the MPE generally still is unclear; thinking about the concept in relation to actual test items allows the judges to be more concrete about how the MPE that they are conceptualizing would handle the test content. In addition, this process provides the opportunity for judges to discuss with their colleagues how they are conceptualizing the MPE and how that translates into their judgments on the particular items. Further, the second round of discussions allows the judges to discuss how they are using the data, whether or not they are incorporating it into their final judgments, and if they are, the extent to which they are doing so.

After judges completed discussing their second round of judgments for the 15-item practice set, they began the same iterative process of reviewing and providing judgments for a set of 75 items. Unlike the 15-item practice sets, the 75-item sets were created with careful attention to content and difficulty parameters so that they represented shortened versions of actual test forms. Judges first reviewed and provided initial judgments for all items without data; they then reviewed data and provided final judgments for each item. For this set of items there was no discussion amongst judges; all of the work was done independently.

The 75-item sets were the same for each of the three independent panels within each of the Step examinations for a given year, and the initial and final Angoff judgments resulting from these 75-item data sets are those that were used for the current analysis.

Data

Following the review of the 75-item set, all initial and final judgments were entered for each judge and were saved into Microsoft Excel worksheets. Individual worksheets were created for each panel within each year for a given examination; this resulted in six distinct Excel workbook files, each of which contained three worksheets. All of the data were read from the Excel spreadsheets into SPSS 14.0 for analysis.

Analysis

Descriptive statistics (mean and SD) for the initial and final (pre- and post-data) cut scores were calculated by year, by panel, and across panels for each examination. In addition, a 2 (initial or final judgment) x 3 (panel) repeated measures analysis of variance (ANOVA) was conducted for each year within each examination to investigate whether any identified changes between pre- and post-data judgments within and across panels were statistically significant. [Note: Recent research has suggested that hierarchical linear modeling (HLM) analysis may be preferable to repeated measures ANOVA when the data meet certain criteria. HLM may be more appropriate in circumstances such as those in which the more strict assumptions of ANOVA cannot be met, when the issue of missing data is a concern, and when it is important to focus more on individual change rather than group differences (Shin, 2009). In the context of the present research, the main question of interest was whether there were mean score changes across groups; for

that reason, the use of analysis of variance was deemed appropriate.] The initial and final cut scores also were graphed in order to provide a clear picture of the pattern of judgments and to provide a straightforward way of visualizing any identifiable patterns in the data. These graphs show individual judges' initial and final cut scores for each panel and across panels within a given year within each examination.

The pre- and post-data cut scores also were applied to cumulative frequency distributions for a reference group of examinees (first-time takers who are students or graduates of U.S. medical schools), and comparisons between the fail rates that resulted from the two sets of judgments were made. These comparisons provide a practical indication of the impact of the different cut scores on a defined population of examinees.

CHAPTER 4: RESULTS

Standards for each of the three USMLE Steps are reviewed approximately every three years. This study reports on data collected from two standard setting reviews for each examination across a period of approximately six years. For each review, three panels were convened. Overall, this structure yielded data from 18 independent standard setting panels. The minimum number of participants in any given panel was eight and the maximum was 12; the majority of panels (12 of 18) had nine or 10 participants. A total of 176 standard setting judges participated in standard setting activities across the 18 panels. Table 1 in Appendix B shows the overall data structure for the study.

There were two main research questions that guided data analysis. The first question asked whether providing performance data resulted in changes in the estimated cut score, and if so, whether there was additional evidence of specific patterns in the data (convergence and systematic shifts in judgments) that would have implications for procedural validity. Evidence of convergence would be a decreased spread of judgments across judges or panels for the post-data judgments in comparison to the pre-data judgments (i.e., decreased standard deviation in judgments after judges reviewed performance data); this pattern can also be identified graphically by plotting the initial and final judgments by judge and panel. Evidence of convergence provides an indication that judges are approaching the task and interpreting the performance information similarly. Alternatively, the finding of a consistent pattern of increasing or decreasing post-data cut scores would be evidence of systematic shifts in judgments; this finding may suggest that judges are using the performance data inappropriately to make more

norm-referenced rather than content-based judgments. Descriptive statistics provided the general results and repeated measures ANOVA examined the statistical significance of the findings.

The second research question was one of practical rather than statistical significance: “What is the practical impact of the resulting cut scores on a defined population of examinees?” To answer this question, the initial and final cut scores were applied to cumulative frequency distributions for a reference group of examinees (first-time takers who are students or graduates of U.S. medical schools) and comparisons between the fail rates that resulted from the two sets of judgments were made.

Research Question 1: Do the estimated cut scores change after performance data are provided and are there patterns in the data indicative of convergence or systematic shifts in judgments?

The first set of analyses investigated: (1) whether or not Angoff judgments changed after judges were given access to examinee performance data; and (2) whether any patterns in the data were indicative of convergence of systematic shifts in judgments. Descriptive statistics (mean cut score and standard deviation for the initial (pre-data) and final (post-data) judgments) and graphical displays of judge- and panel-level initial and final judgments provide an overall picture of how cut scores changed between initial and final judgments, and results of the repeated measures ANOVAs provide an indication of the statistical significance of the findings.

Tables 4.1, 4.2 and 4.3 present descriptive statistics (mean and standard deviation) for the initial and final cut scores by year and by panel for the Step 1, Step 2 and Step 3 examinations, respectively. These results provide some insight into the question of whether there are systematic shifts in judgments following the review of performance data: For six of the 18 total panels across the three examinations, the cut score decreased after receiving performance data (Step 2 Year 1 Panel 3 (-.78), Step 2 Year 2 Panel 3 (-.83), Step 3 Year 1 Panel 2 (-.82), Step 3 Year 2 Panels 1 (-4.68), 2 (-1.51), and 3 (-1.04); for the remaining 12 panels, final cut scores were higher than initial cut scores. Cut scores increased following the review of performance

Table 4.1. Step 1 Panel-level Descriptive Statistics for Initial and Final Judgments

Standard Setting Year	Panel	Judgment	Mean Cut Score	Standard Deviation
1	1	Initial	61.19	17.60
		Final	63.61	17.01
	2	Initial	61.53	19.22
		Final	62.21	18.05
	3	Initial	59.09	18.57
		Final	60.69	18.20
2	1	Initial	54.61	16.04
		Final	58.71	12.72
	2	Initial	56.74	20.44
		Final	60.58	15.88
	3	Initial	51.81	20.40
		Final	61.87	13.86

data for all of the Step 1 panels across both standard setting years; though the other two examinations did not have a pattern that was consistent across both years, the Step 3 Year 2 data did display a decrease in cut scores following the review of performance data for all three panels. For all examinations and all years, the change in mean cut scores between the initial and final judgments was statistically significant ($p < .05$). In addition, with the exception of Step 1 Year 1 ($F=1.57, p = .21$) there was a significant interaction ($p < .05$) between mean initial and final cut scores by panel, indicating that the way that cut

Table 4.2: Step 2 Panel-level Descriptive Statistics for Initial and Final Judgments

Standard Setting Year	Panel	Judgment	Mean Cut Score	Standard Deviation
1	1	Initial	60.19	19.59
		Final	61.17	16.45
	2	Initial	55.19	20.63
		Final	59.23	18.64
	3	Initial	65.92	19.71
		Final	65.14	14.74
2	1	Initial	64.89	18.28
		Final	68.22	15.00
	2	Initial	63.91	17.84
		Final	64.50	15.91
	3	Initial	67.35	20.40
		Final	66.52	16.94

scores changed between initial and final judgments (i.e., whether they increased or decreased) varied across the individual panels.

Descriptive statistics also were computed across panels for each year and examination by calculating the mean and standard deviation of the panel means. For all but one of the six standard setting years (Step 3 Year 2), the overall mean cut score increased following review of performance information. These results are included as Table 2 in Appendix B.

The issue of convergence was evaluated by reviewing the standard deviation results as well as graphical displays of pre- and post-data judgments. The results indicate a post-data decrease in standard deviation for all 18 panels across all six of the standard

Table 4.3: Step 3 Panel-level Descriptive Statistics for Initial and Final Judgments

Standard Setting Year	Panel	Judgment	Mean Cut Score	Standard Deviation
1	1	Initial	61.61	16.46
		Final	64.91	13.38
	2	Initial	71.89	14.42
		Final	71.07	11.96
	3	Initial	62.78	16.54
		Final	64.66	12.31
2	1	Initial	75.23	14.23
		Final	70.55	12.95
	2	Initial	75.35	15.82
		Final	73.84	14.17
	3	Initial	68.19	15.84
		Final	67.15	13.26

setting meetings; this decrease ranged from .37 (Step 1 Year 1 Panel 3) to 6.54 (Step 1 Year 2 Panel 3). The chance of all 18 standard deviations changing in the same direction by chance may be viewed as $1/2^{17}$, so the result is clearly statistically significant. The variability of the panel means within each of the six standard settings is slightly different; the standard deviation for the final cut score actually increased slightly for two of the standard setting years: for Step 1 Year 1 the standard deviation increased from 1.33 to 1.46 and for Step 2 Year 2 the standard deviation increased from 1.78 to 1.86. For the other years, the average decrease in standard deviation was 1.5 and ranged from .75 (Step 3 Year 2) to 2.35 (Step 2 Year 1). (These results are included as Table 2 in Appendix B.)

Graphical display of the mean initial and final cut scores provides a visual representation of the direction and magnitude of the cut score differences between initial and final judgments and can provide insight into not only *if* the judges are changing their cut scores following the receipt of performance data, but if they are, *how* they are changing. Looking at the data graphically can provide a quick picture of whether judges appear to always follow the data in the same way and either increase or decrease their cut scores as a result, whether there is evidence of convergence (i.e., the spread of cut scores decreases following review of data), and/or whether the cut scores tend to remain the same after reviewing the data.

Overall, review of the graphical displays indicates that for all panels for all examinations, judges for the most part made relatively minor adjustments to their judgments following the receipt of performance data. The graphical information does not suggest that judges follow the data in the same way, as there were both increases and

decreases in final cut scores for almost all panels. In addition, the overall graphical results suggest that convergence is more common than a lack of convergence: 10 panels displayed what might be considered a clear pattern of convergence based on the changes between initial and final judgments. The Step 1 and Step 2 results were fairly similar across panels and standard setting years: the majority of judges had minor cut score changes between initial and final judgments and an overall picture of convergence; only four of the 12 panels did not have a pattern that was readily identifiable simply from reviewing the results graphically. For Step 3, four of the six panels had a clear picture of convergence, while one panel from each year did not.

Figures 4.1 and 4.2 present examples of the graphical results to illustrate the kinds of patterns that were present in the data; figures for all panels are included in Appendix C. Figure 4.1 presents mean initial and final cut scores for Step 1 Year 1 Panel 1; each line in each graph represents an individual judge. (Please note that the scale differs across most examinations and is based on the actual spread of the data.) This graph shows that, for the most part, judges made minor adjustments to their cut scores after reviewing performance data; a notable exception is the one judge (represented by the solid line) who changed from an initial judgment of approximately 48 to a final judgment of approximately 70. Overall, some judges increased and some decreased their cut scores following receipt of data, and the range of mean cut scores appears to have decreased slightly between the initial and final judgments which is indicative of convergence.

Figure 4.1. Step 1 Year 1 Panel 1 Mean Cut Scores for Initial and Final Judgments

Step 2 Year 1 Panel 2 results are presented in Figure 4.2; this graph provides a picture that is somewhat different than most of the other panel-level results because it is not clear simply from looking at the graph whether the judgments converged following the review of data. Of particular interest in that graph is that, while most judges made moderate post-data changes, two of the judges had initial judgments that were relatively close and then following the review of data changed their cut scores fairly substantially and in opposite directions. The descriptive results indicate a decrease in standard deviation between

initial and final judgments and therefore that there was some convergence following the review of data, but examination of the graph alone would not have provided a clear indication of that result.

Figure 4.2. Step 2 Year 1 Panel 2 Mean Cut Scores for Initial and Final Judgments

The next set of results provides a graphical display of the changes between initial and final cut scores at the panel level. For all of the Steps and for all of the panels within

a Step, the change between the initial and final cut scores was relatively small; the figures do indicate that the type of change (i.e., whether cut scores increased or decreased after receiving data) and the extent of the change varied across panels. For Step 1, final judgments were higher than initial judgments for all panels and the picture of convergence was clear only for Year 2. For Step 2, four of the six panels had higher final judgments; one panel each year had lower final judgments, and the picture of convergence was clear only for Year 1. As shown in Figure 4.3, the three Year 1 panels started out with initial judgments ranging from 55.19 (Panel 2) to 65.92 (Panel 3); following receipt of performance data, the post-data judgments for Panels 1 and 2 increased and those for Panel 3 decreased. These changes resulted in a range of post-data judgments from 59.23 (Panel 2) to 65.14 (Panel 3). The clarity of this graphical representation is reinforced by the descriptive results that indicated that the decrease in standard deviation was the largest for this standard setting year. The Step 3 Year 1 results also provide an indication that the panels converged following receipt of the performance data. As can be seen in Figure 4.4, the Year 2 results are less clear; the cut scores for all three panels decreased between initial and final judgments, but even though the standard deviation decreased, the range in the cut scores was relatively unchanged.

Figure 4.3: Step 2 Year 2 Panel-level Mean Scores for Initial and Final Judgments

Research Question 2: What is the practical impact of the resulting cut scores on a defined examinee population?

The final set of analyses investigated the practical implications of the ANOVA results by applying the pre-and post-data cut scores to cumulative frequency distributions for a reference group of examinees (first-time takers who are students or graduates of U.S. medical schools). Comparisons between resulting fail rates from the two sets of judgments provide a practical indication of the impact of the cut scores on a defined

population of examinees. All fail rate information can be found in Tables 9-12 in Appendix B.

Figure 4.4. Step 3 Year 2 Panel-level Mean Scores for Initial and Final Judgments

Review of the Step 1 results for both standard setting years indicates that for all but one of the panels the percent of failing examinees increased after judges reviewed performance data; the smallest change was a .8% increase for Year 1 Panel 3 (from 3.4% to 4.2%), and the largest change was a 6.6% increase for Year 2 Panel 3 (from 1.3% to 7.9%). The one panel that was the exception was Year 1 Panel 2; though there was a slight increase in the mean cut score for final judgments, both the initial and final cut scores would have failed 5.5% of examinees.

The Step 2 results present a slightly different picture: for this examination, review of the data led to an increase in the percent of failing examinees for three panels, a decrease in the percent of failing examinees for two of the panels, and no change in the percent of failing examinees for one panel. The smallest increase was a change of 2.7% for Year 2 Panel 2, and the largest change was an 8.1% increase in failing examinees in Year 2 Panel 1. For the two panels that displayed lower percents of failing examinees following review of the data, one decreased by 3.5% (Year 1 Panel 3) and one by 2.6% (Year 2 Panel 3). While the cut score increased slightly after review of data in Year 1 Panel 1, the change (from 60.19 to 61.17) did not result in any change to the 4.2% fail rate.

For Step 3, the results were quite different from those of the other two panels. The fail rate increased for two panels (Year 1 Panels 1 and 3) and stayed the same for one panel (Year 1 Panel 2) following review of the data. The two panels for which the fail rates increased had failing percents that were among the lowest across all examinations (1.2% and 2.0% before data and 3.3% for both following data, respectively), and the

increases were similarly small, increasing 1.3% for Panel 3 and 2.1% for Panel 1). The most notable finding was that all three panels in Year 2 had decreased fail rates following review of data. In both Panel 1 and Panel 2, the fail rate resulting from the initial judgments was 58.5%; this fail rate dropped to 22.8% and 48.5% for Panels 1 and 2, respectively, following the review of data. Panel 3 displayed a more modest initial fail rate of 16.6% which decreased to 11.3% following the review of data.

Table 4.4 presents the fail rate results collapsed across panels within each standard setting for each examination. For four of the six standard settings, judgments made following the review of performance data resulted in a higher fail rate than did those made in the absence of data; these changes ranged from an increase of 1.3% for Step 1 Year 1 to an increase of 4.3% for Step 1 Year 2. Only one standard setting had a final percentage of examinees failing that was the same as what resulted from the initial judgments: Step 2 Year 2 judgments would have failed 8.8% of examinees regardless of whether or not the judges reviewed examinee performance data. Finally, the Step 3 results for Year 2 are the only ones in which review of the performance data led to a decrease in the percentage of examinees failing from 39.1% to 22.8%. While the changes in fail rate may seem modest in many cases, fail rates are generally low overall; even seemingly modest changes between initial and final judgments are therefore likely to represent practically significant increases in the percent of examinees failing.

Table 4.4. Mean Initial and Final Cut Scores and Associated Percents of Failing Examinees for All Examinations and Standard Setting Years

	Standard Setting Year	Judgment	Mean Cut Score	Percent of Failing Examinees
Step 1	1	Initial	60.60	4.2
		Final	62.17	5.5
	2	Initial	54.38	2.1
		Final	60.38	6.4
Step 2	1	Initial	60.43	5.1
		Final	61.85	6.9
	2	Initial	65.38	8.8
		Final	66.41	8.8
Step 3	1	Initial	65.43	5.1
		Final	66.88	7.8
	2	Initial	72.92	39.1
		Final	70.51	22.8

CHAPTER 5: DISCUSSION

The present research provided a large-scale investigation of the impact of examinee performance information on the outcomes of modified Angoff standard setting exercises. Some researchers have questioned the practice of providing performance data, suggesting that resulting cut scores are less likely to be impacted than is panelist variability (Hambleton, 2001). Other research has found significant cut score changes for a majority of studies after judges reviewed data (Brandon, 2004). Results of the present study, which included 18 replications of a standard setting procedure across the three USMLE multiple-choice examinations, indicate that both panelist variability and resulting cut scores were affected by the data. In general, panelist variability decreased after judges reviewed performance data. In addition, post-data cut scores were significantly different from those in the pre-data condition; these differences were observed for all panels across all examinations. Fail rate changes were associated with the cut score changes for a majority of standard setting years and provide additional evidence of the practical significance of the results.

A follow-up to the question of whether cut scores changed is one of *how* they changed and whether it was in a systematic or predictable way. Results of studies dedicated to investigating this topic vary; one meta-analysis concluded that cut scores following data review generally were lower than those from the pre-data condition (Hurtz and Auerbach, 2003). The present results did not suggest the presence of a systematic effect, as the direction of the cut score change varied across panels: for 12 panels the final cut scores were higher than the initial cut scores, and for the remaining six panels the cut

scores decreased after data review. Cut scores consistently increased across all of the Step 1 panels, while for the other two examinations a systematic pattern of increasing or decreasing final cut scores was not found. One factor that may contribute to the systematic change for the Step 1 panels has to do with the realities of the Step 1 testing scenario. Step 1 is a test of basic science knowledge and is most commonly taken by United States medical students in or around their second year of study. The panelists who participate in standard setting activities are practicing physicians, and they have not studied the specific test content nor taken Step 1 or its equivalent for a minimum of five years (and, as is more realistically the case, for a decade or more). As a result, they often find the test material somewhat difficult, and this leads them to overestimate the difficulty for minimally proficient examinees. When they review the data, it is often the case that a much higher proportion of all examinees than expected get the items correct. This is for a number of reasons. First, as opposed to the panelists who have not studied basic science in years, Step 1 examinees spent the two years prior to taking the examination studying the specific material that is covered on the examination. In addition, they also spend several months specifically preparing for the Step 1 examination itself. As a result, their knowledge of the material will never be more extensive than it is at that point. When panelists are reminded of the timing and test preparation issues, it generally results in an upward adjustment to their estimates. Test preparation also occurs for the other Steps, but it is most intensive for Step 1. In addition to the temporal disconnect experienced by panelists, the content of the Step 1 examination (vs. Steps 2 and 3) has the least direct relevance to actual practice. This

disconnect between the test content and experience may make the material seem more difficult, and as a result judges tend to provide lower initial estimates that they adjust upward following the review of data.

One aspect of the examinations that were studied in this research that may differ from other standard setting contexts is that current fail rates are public knowledge. Most panelists already know the current fail rates for the US medical student population; when they don't, the question is almost always asked during the course of the initial training. While it is fine for judges to know that information, it should be irrelevant for the purposes of the content-based standard setting exercises. Their role as judges is to review test content and to make decisions about that content based on expectations for minimally proficient examinees; there is no expectation that new standards will be similar to old standards, and in fact the judges' input would not be necessary if the desire was to maintain existing fail rates. Despite the explicit cautions, some panelists may have leaned on the data to make decisions that were in line with their expectations or with existing statistics. This clearly would not be a concern in new examination contexts (without historical fail rates) or in contexts where there is no expectation that fail rates will be similar from one administration to the next. In the present context, however, the extent to which judges allowed the knowledge of current fail rates to influence their judgments directly impacts the extent to which the judgments were truly content based (as intended) rather than norm referenced.

The present research involved one step in the process of collecting data to inform cut score decisions, but cut scores resulting from this process are not directly used to

make decisions about examinees. Instead, decisions about whether a cut score change is warranted and what that change will be are made by examination-specific oversight groups. The Angoff data are an important source of information used by the policymakers to inform their decisions, but the cut scores resulting from the panels are not used directly. As such, the practical impact of the use of performance data may be more limited in this context than in other contexts where the results of Angoff panels are directly translated into the new cut scores. It is especially important for practitioners in these contexts to carefully evaluate the procedures that they use to generate cut scores in order to be confident that the score are a valid reflection of expert judgment.

One of the methodological strengths of the present research is that there were multiple replications of the same process across different examinations. This aspect of the research design lends some confidence to the resulting conclusions, but the fact remains that only one specific standard setting method was studied and only one procedure that included two specific types of performance data was used. Making broad generalizations about the present findings, particularly in the context of differences in standard setting procedures or methodologies, therefore should be done with caution.

The *Decision* component of Kane's (2006) argument-based approach to validity is concerned with evidence that the process used to establish cut scores and make decisions about examinees was reasonable. Two distinct findings from the present research provide evidence for procedural validity of the USMLE content-based standard setting process. First, convergence of judgments within panels provides some evidence that post-data changes were not random and instead that judges interpreted and used the data similarly

when revising their initial judgments. Second, the absence of systematic changes lends some confidence to the interpretation that resulting cut scores are a reflection of the judges' expectations based on test content and are not simply a reflection of the norm-referenced data they reviewed. Both of these findings provide important insight into the process and important evidence to support decisions and inferences that are made based on the resulting cut scores.

Existing research on the topic of procedural modifications in Angoff-style standard setting exercises is plentiful but has been somewhat limited in scope, generally reporting on results from a small number of replications and a limited number of panels; this limited scope impacts the ability to make overall interpretations about the findings. The structure of the present study, which included 18 replications of the same process across three different examinations, provides the most robust evaluation of the practice of providing performance data that has been reported in the literature. Results of this large-scale investigation make it clear that providing standard setting judges with performance data makes a difference. There was nothing systematic about the findings, and by extension there was nothing that called into question the validity of the standard setting process. As a result, score users can maintain confidence in this procedure and the resulting decisions and interpretations that are made about examinees based on the outcomes. Collecting evidence for the validity of standard setting procedures is an important step in the overall process of developing a validity argument for the USMLE examination system; this ongoing work will help score users feel confident that the scores

that result from the testing process are valid reflections of proficiencies the tests are intended to measure.

The implications of this research extend beyond the specific testing context in which the work was done. Practitioners involved in the standard setting process - regardless of whether a test is being used to make decisions about the level of algebra proficiency of 8th grade students or competence to practice law- should be able to provide evidence that the procedures that were used were reasonable; this evidence provides score users with some confidence in the outcomes and, by extension, in the interpretations that will be made about test takers. It is hoped that the present results will provide insight into the practice of providing performance data during Angoff standard setting exercises and will indicate to practitioners that the specifics of how the standard setting procedure is conducted will have a direct impact on the validity of the results.

In the specific credentialing context that was the focus of the present research, outcomes of the testing process have a direct impact on the public. Test takers who receive passing scores become eligible to receive a license to practice medicine, and it therefore is the responsibility of the testing organization to do everything within its power to ensure that decisions about who passes (and who fails) the examination are the result of sound and defensible procedures. An examinee who passes the USMLE system and is eligible to receive a license to practice medicine should pass because s/he has enough of the necessary content knowledge to provide safe and effective patient care. Collecting information about the procedures that are used to inform cut score decisions will help to ensure that the decisions are based on the required level of examinee proficiency rather

than a result of a procedural flaw; this will, by extension, provide evidence that resulting decisions made about examinees are reasonable. Collecting information of this sort, particularly when the testing outcomes have a significant impact on the well-being of society as a whole, is a critical part of the overall testing process. According to Lee Cronbach: “A proposition deserves some degree of trust only when it has survived serious attempts to falsify it.” (Cronbach, 1980, p. 3). Only after asking difficult questions and being critical of current processes can practitioners and the public have confidence in the outcomes.

REFERENCES CITED

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.)* (pp. 508-600). Washington, DC: American Council on Education.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137-172.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59-88
- 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.
- Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice*, 15(2), 20-31.
- Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement*, 46, 390-407.

- Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). Empirical evidence for the evaluation of performance standards estimated using the Angoff procedure. *Applied Measurement in Education*.
- Clauser, B. E., Swanson, D. B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement*, 39,269-290.
- Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement over a decade* (Vol 5). San Francisco: Jossey-Bass.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teachers Examination. *Journal of Educational Measurement*, 21, 113-129.
- Ebel, R. L. (1972). *Essentials of educational measurement*. (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Fehrmann, M. L., Woehr, D. J., & Arthur, W. Jr. (1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. *Educational and Psychological Measurement*, 51, 857-872.

- Flanagan, J. C. (1948). The aviation psychology program in the Army Air Forces.
Washington: U.S. Government Printing Office.
- Ford, H. (2010). Norm-referenced testing. Retrieved June 10, 2010 from
<http://www.education.com/reference/article/norm-referenced-testing>.
- Goodwin, L.D. (1999). Relations between observed item difficulty levels and Angoff
minimum passing levels for a group of borderline candidates. *Applied
Measurement in Education*, 12(1), 13-28.
- Hambleton, R. K. (2001). Setting performance standards on educational assessments
and criteria for evaluating the process. In G. Cizek (Ed.) *Standard setting:
Concepts, methods and perspectives* (pp. 159-173). Mahwah, NJ: Erlbaum
Associates.
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000). Setting
performance standards on complex educational assessments. *Applied
Measurement in Education*, 24, 355-366.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L.
Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport,
CT: American Council on Education/Praeger.
- 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.

- Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). Englewood Cliffs, NJ: Prentice Hall.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461.
- Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment*, 5(3), 129-145.
- Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.) *Standard setting: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum Associates
- Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed). Westport, CT: American Council on Education/Praeger.
- Livingston, S. & Zieky, M. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: ETS.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education, MacMillan Publishing Co.
- Mills, C. N. (1995). Establishing passing standards, In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 219-252) Lincoln, NE: Buros Institute of Mental Measurements.

- Nedelsky, (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.
- Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement*, 25(1), 57-65.
- Plake, B. S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice*, 27(1), 3-9.
- Plake, B. S., Impara, J., & Potenza, M. (1994). Content specificity of expert judges in a standard setting study. *Journal of Educational Measurement*, 31, 339-347.
- Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, 4, 447-467.
- Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.) *A guide to criterion referenced test construction* (pp. 169-198). Baltimore: Johns Hopkins University Press.
- Shepard, L. A. (1995). Implications for standard setting of the National Academy of Educational Evaluation of the National Assessment of Educational Progress achievement levels. In *Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board and the National Center for Educational Statistics*. (pp. 143-159). Washington, DC: U.S. Government Printing Office.
- Shin, J.H. (2009). Application of repeated-measures analysis of variance and hierarchical linear model in nursing research. *Nursing Research*, 58(3), 211-217.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issues, illustrations, and recommendations. *Human Performance, 9*, 275-295.

APPENDIX A: Sample Performance Data Presented to Standard Setting Judges

Options	A	B	C	D	E
Percent Selected	5	4	82	8	0

APPENDIX B: TABLES

Table B1. Summary of Overall Dataset Structure

Examination	Standard Setting Year	Panel	N
Step 1	Year 1: 2003	1	10
		2	9
		3	10
	Year 2: 2006	1	9
		2	10
		3	10
Step 2	Year 1: 2002	1	10
		2	8
		3	9
	Year 2: 2007	1	9
		2	8
		3	11
Step 3	Year 1: 2004	1	9
		2	9
		3	10
	Year 2: 2008	1	12
		2	12
		3	11
Totals	6	18	176

Table B2. Mean Cut Scores for Each Standard Setting Year Calculated by Taking the Mean of the Individual Panel Means (N=3)

	Standard Setting Year	Judgment	Mean Cut Score	Standard Deviation
Step 1	1	Initial	60.60	1.33
		Final	62.17	1.46
	2	Initial	54.38	2.47
		Final	60.38	1.59
Step 2	1	Initial	60.43	5.37
		Final	61.85	3.02
	2	Initial	65.38	1.78
		Final	66.41	1.86
Step 3	1	Initial	65.43	5.63
		Final	66.88	3.63
	2	Initial	72.92	4.10
		Final	70.51	3.35

Table B3a. Step 1 Year 1 Repeated Measures ANOVA Results: Tests of Within-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Judgment (Initial, Final)	2646.71	1	2646.71	15.46	.000
Judgment by Panel	538.29	2	269.14	1.57	.208
Error (Judgment)	371804.26	2172	171.18		

Table B3b. Step 1 Year 1 Repeated Measures ANOVA Results: Tests of Between-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Intercept	8175585.89	1	8175585.89	33769.13	.000
Panel	2614.38	2	1307.19	5.40	.005
Error	525846.39	2172	242.10		

Table B4a: Step 1 Year 2 Repeated Measures ANOVA Results: Test of Within-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Judgment (Initial, Final)	39157.60	1	39157.60	288.34	.000
Judgment by Panel	9101.52	2	4550.76	33.51	.000
Error (Judgment)	294696.57	2170	135.81		

Table B4b. Step 1 Year 2 Repeated Measures ANOVA Results: Test of Between-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Intercept	7135388.66	1	7135388.66	32806.32	.000
Panel	1816.45	2	908.23	4.176	.015
Error	471975.90	2170	217.50		

Table B5a: Step 2 Year 1 Repeated Measures ANOVA Results: Test of Within-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Judgment (Initial, Final)	1984.03	1	1984.03	13.06	.000
Judgment by Panel	3736.73	2	1868.37	12.30	.000
Error (Judgment)	304487.98	2004	151.94		

Table B5b. Step 2 Year 1 Repeated Measures ANOVA Results: Test of Between-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Intercept	7451455.63	1	7451455.63	28572.81	.000
Panel	22378.43	2	11189.21	42.91	.000
Error	522619.78	2004	260.79		

Table B6a: Step 2 Year 2 Repeated Measures ANOVA Results: Test of Within-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Judgment (Initial, Final)	1093.77	1	1093.77	6.53	.011
Judgment by Panel	3257.50	2	1628.75	9.73	.000
Error (Judgment)	351197.44	2097	167.48		

Table B6b. Step 2 Year 2 Repeated Measures ANOVA Results: Test of Between-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Intercept	8963541.68	1	8963541.68	39675.79	.000
Panel	2870.59	2	1435.29	6.35	.002
Error	473753.58	2097	225.92		

Table B7a: Step 3 Year 1 Repeated Measures ANOVA Results: Test of Within-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Judgment (Initial, Final)	2210.82	1	2210.82	29.09	.000
Judgment by Panel	2959.28	2	1479.64	19.47	.000
Error (Judgment)	159392.15	2097	76.01		

Table B7b. Step 3 Year 1 Repeated Measures ANOVA Results: Test of Between-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Intercept	9167134.30	1	9167134.30	55022.73	.000
Panel	29250.91	2	14625.45	87.79	.000
Error	349373.47	2097	166.61		

Table B8a: Step 3 Year 2 Repeated Measures ANOVA Results: Test of Within-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Judgment (Initial, Final)	6721.69	1	6721.69	85.39	.000
Judgment by Panel	3250.53	2	1625.27	20.65	.000
Error (Judgment)	183815.25	2335	78.72		

Table B8b. Step 3 Year 2 Repeated Measures ANOVA Results: Test of Between-subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Significance
Intercept	11893165.51	1	11893165.51	71835.12	.000
Panel	18988.24	2	9494.12	57.35	.000
Error	386587.23	2335	165.56		

Table B9. Step 1 Initial and Final Cut Scores and Associated Percent Failing by Panel

		Judgment	Mean Cut Score	Percent of Failing Examinees
2003	1	Initial	61.19	4.2
		Final	63.61	6.8
	2	Initial	61.53	5.5
		Final	62.21	5.5
	3	Initial	59.09	3.4
		Final	60.69	4.2
2006	1	Initial	54.61	2.1
		Final	58.71	5.2
	2	Initial	56.74	3.5
		Final	60.58	6.4
	3	Initial	51.81	1.3
		Final	61.87	7.9

Table B10. Step 2 Initial and Final Cut Scores and Associated Percent Failing by Panel

Standard Setting Year	Panel	Judgment	Mean Cut Score	Percent of Failing Examinees
2002	1	Initial	60.19	4.2
		Final	61.17	4.2
	2	Initial	55.19	1.6
		Final	59.23	3.4
	3	Initial	65.92	10.3
		Final	65.14	6.8
2007	1	Initial	64.89	6.5
		Final	68.22	14.6
	2	Initial	63.91	4.8
		Final	64.50	6.5
	3	Initial	67.35	11.4
		Final	66.52	8.8

Table B11. Step 3 Initial and Final Cut Scores and Associated Percent Failing by Panel

Standard Setting Year	Panel	Judgment	Mean Cut Score	Percent of Failing Examinees
2004	1	Initial	61.61	1.2
		Final	64.91	3.3
	2	Initial	71.89	21.7
		Final	71.07	21.7
	3	Initial	62.78	2.0
		Final	64.66	3.3
2008	1	Initial	75.23	58.5
		Final	70.55	22.8
	2	Initial	75.35	58.5
		Final	73.84	48.5
	3	Initial	68.19	16.6
		Final	67.15	11.3

Table B12. Mean Initial and Final Cut Scores and Associated Percents of Failing Examinees for All Examinations and Standard Setting Years

	Standard Setting Year	Judgment	Mean Cut Score	Percent of Failing Examinees
Step 1	1	Initial	60.60	4.2
		Final	62.17	5.5
	2	Initial	54.38	2.1
		Final	60.38	6.4
Step 2	1	Initial	60.43	5.1
		Final	61.85	6.9
	2	Initial	65.38	8.8
		Final	66.41	8.8
Step 3	1	Initial	65.43	5.1
		Final	66.88	7.8
	2	Initial	72.92	39.1
		Final	70.51	22.8

APPENDIX C: FIGURES

Figure C1. Step 1 Year 1 Panel 1 Individual Judge Initial and Final Cut Scores

Figure C2. Step 1 Year 1 Panel 2 Individual Judge Initial and Final Cut Scores

Figure C3. Step 1 Year 1 Panel 3 Individual Judge Initial and Final Cut Scores

Figure C4. Step 1 Year 2 Panel 1 Individual Judge Initial and Final Cut Scores

Figure C5. Step 1 Year 2 Panel 2 Individual Judge Initial and Final Cut Scores

Figure C6. Step 1 Year 2 Panel 3 Individual Judge Initial and Final Cut Scores

Figure C7. Step 2 Year 1 Panel 1 Individual Judge Initial and Final Cut Scores

Figure C8. Step 2 Year 1 Panel 2 Individual Judge Initial and Final Cut Scores

Figure C9. Step 2 Year 1 Panel 3 Individual Judge Initial and Final Cut Scores

Figure C10. Step 2 Year 2 Panel 1 Individual Judge Initial and Final Cut Scores

Figure C11. Step 2 Year 2 Panel 2 Individual Judge Initial and Final Cut Scores

Figure C12. Step 2 Year 2 Panel 3 Individual Judge Initial and Final Cut Scores

Figure C13. Step 3 Year 1 Panel 1 Individual Judge Initial and Final Cut Scores

Figure C14. Step 3 Year 1 Panel 2 Individual Judge Initial and Final Cut Scores

Figure C15. Step 3 Year 1 Panel 3 Individual Judge Initial and Final Cut Scores

Figure C16. Step 3 Year 2 Panel 1 Individual Judge Initial and Final Cut Scores

Figure C17. Step 3 Year 2 Panel 2 Individual Judge Initial and Final Cut Scores

Figure C18. Step 3 Year 2 Panel 3 Individual Judge Initial and Final Cut Scores

Figure C19. Step 1 Year 1 Panel-level Initial and Final Cut Scores

Figure C20. Step 1 Year 2 Panel-level Initial and Final Cut Scores

Figure C21. Step 2 Year 1 Panel-level Initial and Final Cut Scores

Figure C22. Step 2 Year 2 Panel-level Initial and Final Cut Scores

Figure C23. Step 3 Year 1 Panel-level Initial and Final Cut Scores

Figure C24. Step 3 Year 2 Panel-level Initial and Final Cut Scores