

FURTER, ROBERT THOMAS, Ph.D. Principled Assessment as a Foundation for Standard Setting (2015)
Directed by Dr. Richard Luecht, 96 pp.

This study investigated the impact of using Assessment Engineering (AE) task models as the unit of judgment in a standard setting workshop. The proposed method, or Task Model-based Standard Setting (TMSS), used a procedure similar to that of the Bookmark Standard Setting Procedure; however, task models, rather than items, served as the unit of judgment. The proposed method was compared against the yes/no Angoff method and the Bookmark Standard Setting Procedure in regard to the rigor of the recommended cut-scores, the interpanelist consistency of the cut-scores, and the panelists' understanding of the score scale. The TMSS was found to produce a similar cut-score in regard to rigor and interpanelist consistency as that of the Bookmark; however, panelists were more comfortable with the TMSS procedure and indicated a better understanding of the score scale than the Bookmark group. Panelists indicated the greatest level of comfort/understanding with the yes/no Angoff procedure, which resulted in the least rigorous cut-score. The Angoff group was found to have the lowest level interpanelist consistency comparatively, and did not align with panelists' holistic. Results in this study indicate using task models to set standards, or at least including them in the process, can facilitate panelist understanding of the scale. Furthermore, the TMSS had the strongest collection of validity evidence of the three methods examined in this study.

PRINCIPLED ASSESSMENT AS A FOUNDATION FOR STANDARD SETTING

by

Robert Thomas Furter

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2015

Approved by

Committee Chair

To my beautiful wife, my loving family, and better friends than any one man deserves.

APPROVAL PAGE

This dissertation written by Robert Thomas Furter has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

Oct 19, 2015

Date of Acceptance by Committee

Oct 19, 2015

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I must take this opportunity to thank my advisor, Dr. Richard Luecht, who is not only a brilliant scholar, but a generous and caring individual. I will be forever indebted to Dr. Luecht and my faculty in the Educational Research Methodology department at the University of North Carolina at Greensboro- Dr. Terry Ackerman, Dr. Chalhoub-Deville, Dr. Holly Downs, Dr. Robert Henson, Dr. Randall Penfield, and Dr. John Willse. Not only did these individuals bestow invaluable knowledge, but more importantly they taught me to analytically and critically pursue knowledge.

I would also like to thank Dr. Scott Arbet, Dr. Matthew Burke, Dr. Mike Clark, and Dr. Stephen Murphy, for their mentorship and substantial contributions to my professional growth. It has truly been a pleasure learning from these individuals, and I look forward to continuing to learn from, and collaborate with, them in the future.

I must thank Dr. Linda Althouse, Dr. Rachael Tan, and Erik Meyer. This dissertation would not have been possible without their support and confidence. I owe them a tremendous debt of gratitude.

Lastly, I would like to acknowledge the incredible contribution my wife, parents, and siblings have made to my academic career. I cannot imagine this journey without their unconditional love and support.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
I. INTRODUCTION	1
II. LITERATURE REVIEW.....	16
The Meaning of Scale Scores.....	16
Principled Assessment Design.....	32
The Validity of Standard Setting.....	37
Conclusion.....	49
III. METHODS.....	50
Task Model-Based Standard Setting.....	50
The Study.....	52
Conclusion.....	64
IV. RESULTS.....	65
Procedural Evidence for the Validity of the Workshops.....	65
Final Cut-Score Rigor and Interpanelist Consistency.....	69
Panelists' Confidence and Ability to Interpret the Scale.....	72
Additional Results: Hofstee ratings.....	75
V. SIGNIFICANCE, LIMITATIONS AND FUTURE RESEARCH	77
REFERENCES	83
APPENDIX A. BOOKMARK AND TMSS RATING SHEETS.....	90
APPENDIX B. SCREENSHOT OF QUALTRICS SURVEY FOR ANGOFF RATINGS.....	92
APPENDIX C. ANGOFF AND BOOKMARK SURVEY.....	93
APPENDIX D. TMSS SURVEY.....	95

LIST OF TABLES

	Page
Table 1. College Board (2013) PLDs for Interpretive Communication at Level 3.....	4
Table 2. Agenda for Standard Setting Groups.....	61
Table 3. Final Survey Response Frequencies.....	66
Table 4. Cut-Scores by Method and Round.....	70
Table 5. Statements Regarding Competency of Passing Candidates.....	74
Table 6. Raw Cut-Scores and Hofstee Acceptable Cut-Scores Ranges.....	75

LIST OF FIGURES

	Page
Figure 1. Example Construct Map and Task Model Map.....	12
Figure 2. Figure 9.5 from Kolen and Brennan (2004, p. 360).....	19
Figure 3. Table 2 from Ebel (1962, p. 20).....	21
Figure 4. Item Difficulty Parameter by Location in OIB.....	54
Figure 5. Distribution of Item Difficulties for Standard Setting Items.....	55
Figure 6. Average Agreement for Survey Responses.....	68
Figure 7. Graphical Display of Cut-Scores by Method and Round	71

CHAPTER I

INTRODUCTION

Many contemporary educational and certification/licensure assessments attempt to group examinees into categories based on the examinees' performance on the assessment. The process of deciding upon appropriate standards, or cut-scores, at which to discern one group of examinees from another is known as standard setting. The process of categorization based on performance serves another purpose in regard to the score scale- it adds meaning and interpretation to the scale. Labels are often attached to the groups created by set standards, such as "pass" or "fail" in credentialing/licensing exams, or "advanced", "proficient", or "basic" in educational achievement testing. Descriptors or claims about the competencies or achievement level associated with the categories usually supplement such labels. In the case of certification/licensure, a candidate who "passed" the exam may be assumed to be at least minimally competent to practice- that is, possess the minimum expected skills and knowledge within a given field with minimal risk to the public. In educational testing, such as an end of grade test, a student who placed in the "proficient" category may be said to have mastered the relevant curriculum within a given subject for his/her grade level. In labeling examinees and making associated inferences as to their knowledge, skills, and abilities (KSAs), we attribute meaning to points on the scale.

The task of attributing meaning to a score scale is often necessary in educational and psychological measurement. The scales that underlie educational assessments generally have little meaning to the public at large, or many test users for that matter, if they are reported as a single number. One method for attributing meaning to a score is norm-referencing; however, norm-referenced scores have serious limitations regarding the inferences that can be made about examinees. Under a norm-referenced approach, we may be able to say that an examinee scores at the 90th percentile; that is, has performed as well or better than 90% of his/her peers. Depending on the purpose and use of this test, knowing an examinee's rank ordering relative to some defined reference group of examinees may suffice. However, if the goal of the assessment is to make claims about an examinee's KSAs, then a criterion-referenced approach is necessary. As Ebel (1962, p. 18) pointed out over 50 years ago, "It is not very useful to know that Johnny is superior to 84 percent of his peers unless we know what it is that he can do better than they, and just how well he can do it!"

Standard setting is one of a few methods by which a scale can be given meaning and interpretability under a criterion-referenced framework. In standard setting, groups of subject matter experts (SMEs) convene and participate in a judgmental process to recommend cut-scores, usually to a governing body or group of policy-makers. In conducting this process of partitioning the scale, panelists, and ultimately policy-makers, are adding qualitative interpretations to the scale. Without such a process to add content and construct-related clarifications to a scale, scale scores tell us little more than which examinees performed better than others. For example, a score of "3" on the Advanced

Placement (AP) French Language and Culture exam does not provide much information unless it is accompanied by the interpretation that a “3” means a student is, “capable of doing the work of an introductory-level course in a particular subject at college” (College Board, 2014). In addition to classifications such as “passing”, “proficient”, or “capable”, the inferences made regarding an examinee’s KSAs are often summarized into descriptive statements in the form of performance level descriptors (PLDs) or achievement level descriptors (ALDs). As defined in Egan, Schneider, and Ferrara (2012, p. 79), “PLDs define the knowledge, skills, and processes (KSPs) of students at specified levels of achievement and often include input from policy makers, stakeholders, and content experts.”

PLDs play two large roles in most contemporary standard setting methods. First, PLDs assist standard setting panels in actually setting the cut-scores, as they help panels typify the KSAs of examinees that constitute a particular ability or proficiency level. Second, once the standard-setting cut scores are obtained, PLDs add qualitative interpretation guidance to the score scale for test users and test takers, allowing for substantive claims to be made regarding candidates KSAs based on their location on the score scale. That is, PLDs allow us to reflect upon the nature and extent of expectations or claims we wish to make regarding examinees scoring at different points on the scale. As an example, Table 1 has been adapted from College Board’s (2013) documentation, and shows the PLDs¹ associated with a score of “3” on the AP French Language and Culture exam. The PLDs displayed below refer specifically to the interpretive mode of

¹ Although used interchangeably in the literature, College Board generally uses the term “achievement level descriptors” (ALDs), rather than “performance level descriptors” (PLDs).

communication, and separate PLDs are available for the interpersonal and presentational modes, respectively.

Table 1. College Board (2013) PLDs for Interpretive Communication at Level 3

	Student receiving a score of 3:
Text Content	Identifies several main ideas and details on familiar topics.
Text Comprehension	Is able to respond accurately to basic information questions (e.g. who, what, when, where-type questions) and can respond to a limited number of questions that require inferring implied meanings.
Text Analysis	Identifies several of the distinguishing features of the text (e.g. the type of resource, intended audience and purpose of the resource, and tone).
Vocabulary	Comprehends a variety of vocabulary from familiar thematic word groups, including some idiomatic expressions.
Context	Is sometimes able to use context to deduce the meaning of unfamiliar vocabulary.
Cultural Awareness & Comparison	Identifies the cultural products and practices of the target culture(s), and demonstrates an understanding of basic content of familiar interdisciplinary topics in the resource. Identifies a few geographic, historical, artistic, social, or political features of target culture communities.

The inferences that we wish to make about test performance are essential aspects of an overall validity argument about any testing program. Messick (1989, p.13) defined validity as, “an integrated judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.” Because standard setting is a key component in the interpretation (i.e., making inferences about an examinee’s KSAs based on a test score) and use (i.e., deciding if someone gets college credit, is fit to practice in a field, etc.) of a test score, it is clear that the process and results of a standard setting procedure have paramount implications for the validity of a test score. The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) acknowledge the connection between test score interpretations and cut-scores in the joint standards (henceforth referred to as the *Standards*) as they state, “...cut scores provide the basis for using and interpreting test results” (2016, p. 2014). Previous editions of the *Standards* also called for evidence to be documented regarding how the standards were set, how panelists were selected for the standard setting procedure, etc., which has resulted in the creation of validity frameworks for standard setting (Kane, 1994; Kane 1999; Norcini & Shea, 1997; Pant, Rupp, Tiffin-Richards, & Koller, 2009). The validity of a standard setting procedure will be addressed in more depth in the following chapter.

Generally speaking, contemporary standard setting methods ask panelists to examine elements of the assessment (exam items or tasks) or samples of examinee work (portfolios or responses to exam items or tasks), and use this information to set cut-

scores. A broad distinction can be made separating standard setting methods into two primary categories- examinee-centered and test-centered methods (Jaeger, 1989; Kane, 1994). In the examinee-centered approaches, panelists categorize samples of examinee performance. The scale locations that correspond to the dividing lines between examinee groups serve as the cut-score locations. In the test-centered approaches, panelists review a sample of test items or tasks. In reviewing the sample of items or tasks, the panelists make judgments as to how examinees are expected to perform on the items or tasks. Some mathematical procedure is then used to combine the expectations across panelists, and within panelists when necessary for some methods.

A host of factors related to the item types involved in an assessment, resources, and other issues play a role in which method a testing program will select (Hambleton & Pitoniak, 2006). As cited in Cizek (2012), Kaftandjieva (2010) estimates the current number of standard setting methods to be greater than 60, but Cizek (2012, p. 9) contends, “If each unique but substantive modification of a basic approach is counted, the number of current methods is essentially infinite.” Regardless of the method employed, standard setting will generally first occur toward the end of the initial testing cycle, after an exam has been administered for the first iteration and scored. Setting standards late in the assessment cycle may seem odd, given the fact that test forms are often built to maximize precision at cut-scores or decision points and the interpretation and use of scores hinges on the cut-scores (as stated by the AERA, APA, and NCME *Standards* [2014]). When data is available following the first administration, cut-scores can be set moving forward to guide subsequent form construction.

The approach of setting standards in an almost “post-hoc” fashion is less surprising when viewed within the context of how the majority of testing programs traditionally operate. As described in Luecht (2013a), content blueprints and test specifications have governed what has become the traditional method for test development and construction. Content blueprints are created by SMEs and consist of topics that represent the construct of interest, along with how much of each topic is to be included on an exam form. Luecht (2013a) describes three primary problems associated with such a system for test construction and development- 1) SMEs may disagree as to specifically what should comprise a content blueprint- or at least how to code particular items with respect to the blueprint categories, 2) content blueprints and statistical test specifications are often developed independent of one another (e.g., SMEs and test developers work largely within the content blueprint space while psychometricians work with estimating the statistical properties of items and the score scale), and 3) most content blue prints or cognitive specifications developed to code the test items have little or no actual association with the statistical properties of those same items. For example, many items codes in exactly the same categories have widely different statistical properties. Arguably, we should require that statistical item properties such as item difficulty or even indicators of “sensitivity” to the underlying trait(s) or “discrimination” can be shown to causally stem from certain cognitive complexity or experimental design factors, The fact that most content and cognitive coding schema used in test blue print do not relate to statistical properties of the test implies that we instead have two rather separate systems

of test specifications that may actually compete with one another and result in less-than optimal test scores.

Luecht's (2013a) solution to the aforementioned issues is known as assessment engineering (AE; Luecht, 2006). AE is a framework that incorporates strong engineering design principles into an integrated development process intended to ensure synergy between the psychometric properties of a score scale and its intended interpretations and performance claims. AE further extends the use of strong engineering principles to the development of an integrated set of item content and cognitive specifications that impact the difficulty and other psychometric properties of families of items. The net result is a unified set of test and item specifications that allow large-scale, ideally low cost production of items that directly inform the intended interpretations along the score scale.

An even broader validity-based framework for test design and development than AE is known as evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003). Both the AE and ECD frameworks focus on clearly delineating the evidentiary claims an assessment attempts to make regarding examinee performance, and creating tasks geared toward assessing said claims. The frameworks share many similarities, and Luecht (2013a) has stated that AE could be seen as a way of implementing ECD.

There are five building blocks within the AE framework. First, construct maps (see Wilson, 2004) outlines the proficiency claims we wish to make regarding examinees at different levels of the score scale, and may be likened to more specific PLDs.

Associated evidence models dictate the evidence/data necessary to support the

proficiency claims of the construct map, and link the proficiency claims to the most relevant AE structure to the current discussion- task models.

Second, task modeling involves the development of grammars or other mechanisms for documenting the specific and precise delineation of cognitive skills, declarative knowledge, contextual features, and auxiliary features that drive the complexity and difficulty of a task. Content is inherent within the task model, but simple content or cognitive coding is not the ultimate goal. Rather, the grain-size and level of specificity in a task model provides a highly detailed and functional specification of the content, desired cognitive levels of skill or knowledge objects, and other experimental task design controls for an entire family of items that inform performance interpretations and proficiency claims at specific points along the intended score scale. Luecht (2006) calls these functional specifications task-model “grammars”- that is, a language for describing the content and cognitive complexity to be built into each item family.

Third, once created from the task model grammar, task models can be organized into a task model map (TMM), with the complexity of a task model governing its location in the TMM along the scale. The locations of the task models can ultimately be verified using empirical data, and instances where discrepancies occur can be examined. The density of the task models within the TMM follows the statistical or psychometric demands for measurement information. However, since the TMM also carries with it the content and cognitive interpretations inherent in the task model grammar, the TMM becomes a singular test development blue print that simultaneously meets the needs of psychometricians and SMEs.

Fourth, the task-model “grammar” and task models constructed from the grammar are used to build detailed item templates to guide item writing and the creation of scoring procedures for all items within each family. At the very least, these templates provide item writers with highly prescriptive guidelines and constraints to ensure that every item produced meets the intended content and cognitive requirements. Somewhat ambitiously, it is even possible in some domains to fully automate the item generation procedures where each task model is instantiated by producing families containing hundreds or even thousands of interchangeable items.

The final engineering step is to statistically verify the statistical operating characteristics of each item family. Undesirably large variation within the item families is managed by either tightening appropriate item production constraints within the templates, better training the item writers, or even modifying the task model grammar and task models to better specify the content, skills, knowledge, information density, etc. associated with those task models.

Creating a scale under an AE framework forces test developers to think in terms of an ordered proficiency scale from the onset of development. As such, an assessment designed under AE offers the potential for *prospective standard setting*, as described by Bejar, Braun, and Tannenbaum (2007). Prospective standard setting is a possibility in this scenario as meaning and interpretation are built into the scale, rather than attributed after the fact. After describing a scale anchoring procedure using SAT® data, Ebel (1962, p. 21) stated, “The second, and possibly more basic way to secure test scores

which content-meaning is to build the meaning into the test, and hence into the test score, by systematic, explicitly specified processes of test construction.”

Below, Figure 1 shows an example of what a construct map and task model map could look like. Along the left side of the figure are the proficiency claims located along the scale. As stated, the proficiency claims on a construct map share similarities with traditional PLDs but should be more specific in comparison. Along the right side of the figure is the task model map, which is the organization of task models along the scale. Each individual circle represents a unique task model, containing the cognitive skills, declarative knowledge, contextual features, and auxiliary features of the task, which prescribe the task, and also serve to locate the task on the scale continuum based on its combination of features. The proficiency claims and the task models are located along the scale continuum, shown in the middle of the figure.

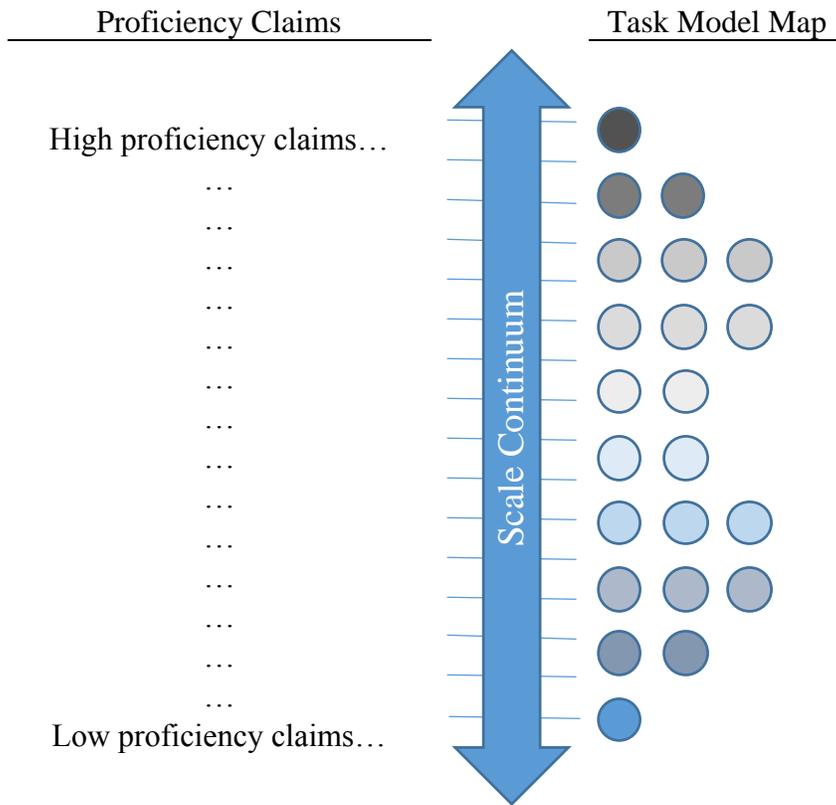


Figure 1. Example Construct Map and Task Model Map

The purpose of this study is to introduce a standard setting method that seeks to incorporate AE task models into the standard setting judgmental task for a large-scale medical certification examination. In incorporating task models into the standard setting judgmental process, the goals are:

- 1) To achieve greater interpanelist consistency of cut-scores,
- 2) To achieve greater panelist understanding of the score scale when compared to other methods,
- 3) And arrive at proficiency claims/expectations of examinees that are better defined and more explicit in comparison to those derived from contemporary standard setting methods.

The first two anticipated outcomes (greater interpanelist consistency and greater panelist engagement) are expected because panelists will be asked to use task models rather than items as their unit of analysis. In a best case scenario, an individual item is a singular instantiation of an ECD or AE task model. Under the predominant testing framework of content blueprints and test specification, an individual item is most likely a unique instantiation of a broad content domain and cognitive level. In using task models, the guess-work and unwanted variability in panelist judgments as to what an item is trying to elicit should be minimized, thereby reducing error and increasing consistency. Similarly, in using task models with specific skill statements, it is expected that panelists will feel a greater connection with the process, more confidence, and a better understanding of how their actions result in the set cut-score. Lastly, the final proficiency claims for passing candidates of the proposed method will be governed by the task models themselves,

providing a stronger link between the proficiency claims and the skills and knowledge an examinee demonstrates on the assessment. If successful, the proposed method could provide for the opportunity of prospective standard setting in the future, allowing for more targeted test construction and a more confirmatory approach to setting performance standards (i.e., test data and criterion measures could serve to validate the standards in ways that are not possible under the current framework).

A standard setting study was conducted to evaluate the three above goals and to simultaneously compare AE-based standard setting with two very prominent, contemporary standard setting methods. The two comparative standard setting methods used for the study were a variation of the Angoff method (Angoff, 1971) and the Bookmark Standard Setting Procedure (Lewis, Green, Mitzel, Baum, & Patz, 1998; Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001). The methods will be compared in regard to their interrater consistency, the panelists' level of engagement and buy-in to each method, and finally the quality and nature of the proficiency claims created by each SME group.

Moving forward, the literature review (Chapter 2) reviews salient literature regarding scaling, standard setting, and principled assessment frameworks, focusing on elements of AE relevant to the proposed method. The methods (chapter 3) will describe the logic behind the proposed standard setting method, and delineate its implementation and evaluation. The methods chapter will also lay out a research study that will compare the proposed method to the two selected contemporary methods. Following the methods chapter, the results (chapter 4) of the standard setting study will be summarized. The dissertation will conclude with a discussion (chapter 5) of the implications and

significance of the findings summarized in the results, the limitations of the study, and potential future research directions.

CHAPTER II

LITERATURE REVIEW

In this chapter, literature related to how meaning is attributed to score scales will be addressed. The discussion will start by briefly covering traditional methods that aim to add interpretability to score scale by relating test content to the scale post-administration, concluding with a treatment of the two contemporary standard setting methods used in this study. The discussion will then move to the principled assessment paradigm of interest for this study, AE, which builds meaning and interpretability into the scale via a system of structures in test development. As this study seeks to incorporate AE into standard setting, this chapter will close with a look at how the validity of standard setting workshops is evaluated. The purpose of addressing the validity of standard setting is to demonstrate how the proposed incorporation of AE structures will adhere to key features of sound practice in standard setting, and where the incorporation of AE structures is anticipated to strengthen the validity argument for the standard setting workshop.

The Meaning of Scale Scores

Kolen and Brennan (2004) and Kolen (2006) describe the process of scaling, along with three methods that have historically been used to incorporate content information and meaning to scale scores. As defined by Kolen and Brennan (2004, p.

329), "...*scaling* is the process of associating numbers or other ordered indicators with the performance of examinees. These numbers and ordered indicators are intended to reflect increasing levels of achievement or ability." A score scale is the result of scaling, and scale scores are used to summarize examinee performance on a measure. Due to the fact that raw scores depend upon the items an examinee saw on a specific form, scale scores are preferred in many cases for equating purposes and to limit confusion that can be created when multiple forms of an assessment are used.

Scaling has been a topic of discussion since the early days of psychometrics. Using Binet intelligence test questions as an example, Thurstone (1925) demonstrated a method for scaling educational and psychological tests. Guttman's (1944) work regarding attitude assessments addressed the importance of ordered scales and the conjunctive nature of ordered scales (i.e., a person located at a given level of a scale can be expected to possess all of the KSAs, or attitudes depending on the purpose of the measure, expressed at lower levels of the scale). Guttman's scaling principles have become synonymous with ordered scaling, as Plake and Cizek (2012, p. 190) used the phrase "Guttman-like properties" in describing hierarchical scales, "with ability systematically increasing as performance moves up the scale..."

Peterson, Kolen, and Hoover (1989, p.222) describe the goal of scaling in stating, "The main purpose of scaling is to aid users in interpreting test results." Historically, adding interpretability to scale scores through the attribution of content information has been done in three primary ways- item mapping, scale anchoring, and standard setting (Kolen & Brennan, 2004). As Kolen and Brennan (2004, p. 358) state, "each of these

methods is intended to help test users understand what examinees who earn particular scale scores know and are able to do.” Kolen and Brennan (2004) discuss these procedures as they would be conducted after the creation of the score scale and administration of an assessment.

Item Mapping

Item mapping involves aligning items to points on the score scale. These items are then shared with test users in an effort to demonstrate the type of items an examinee at a particular score level would answer correctly. To align items to points on the score scale, a response probability (RP) must first be decided that is associated with mastery on all items of the assessment, expressed as a percentage. Kolen and Brennan (2004) give an example using NAEP data, in which an RP of 74% was selected as demonstrating mastery. The RP is then used as the basis for locating items on the score scale by calculating the score scale point associated with an expected probability of a correct response equal to the RP. In regard to the NAEP example, this means that items were aligned to the scale where an examinee with a given scale score would have a 74% probability of responding correctly to the item.

Two criteria often used for selecting items that will be mapped onto the scale are item discrimination and item content domain. Item discrimination is used to ensure that items selected for the map are adequately discriminating between ability levels. The item content domain can be helpful in selection to promote adequate content representation in the item map, as not all items from an assessment may be included on the item map. Similarly, Kolen and Brennan (2004) point out actual items may not always be used on

an item map, and statements summarizing the items may be used instead. Below, Figure 9.5 from Kolen and Brennan (2004) gives an example of an item map that uses statements rather than actual items.

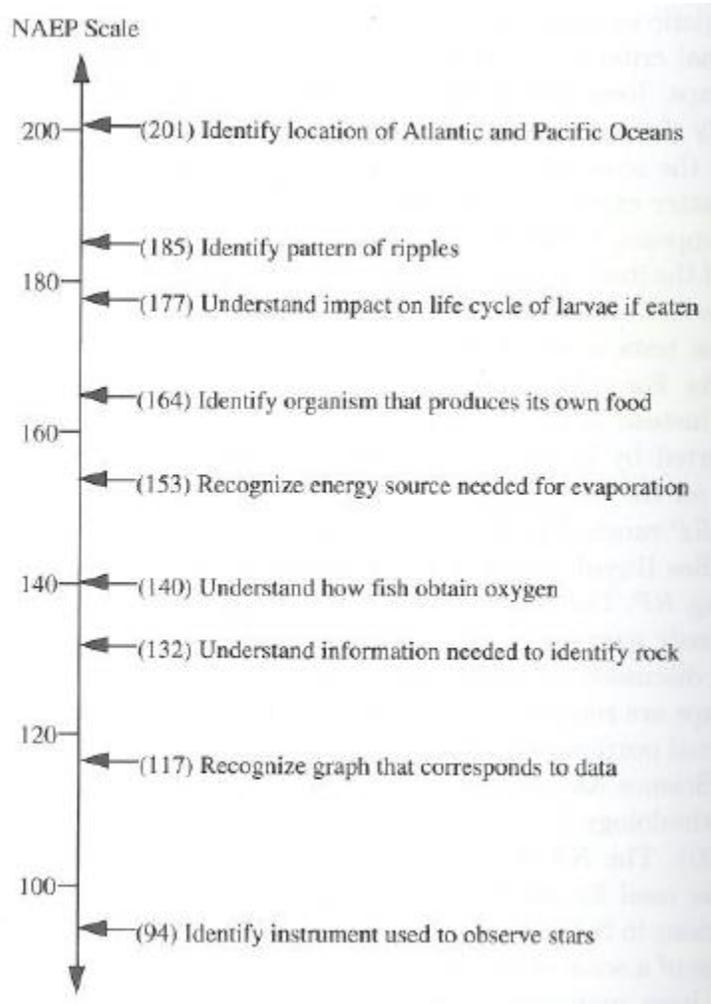


Figure 2. Figure 9.5 from Kolen and Brennan (2004, p. 360)

Scale Anchoring

Scale anchoring is a similar process to that of item mapping; however, in scale anchoring, general statements regarding examinee KSAs are used rather than specific items. Scale anchoring generally begins by selecting scale points that are either equally spaced throughout the scale or are common percentile values. Kolen and Brennan (2004) give an example of 10th, 25th, 50th, 75th, and 90th percentile values. Item maps are then created, consisting of items that are located near the selected scale points and discriminate well at the selected scale points. Following the creation of the item maps, SMEs review the items at each location to generate statements in regard to the KSAs expected of examinees at each of the selected scale points. Kolen and Brennan (2004, p.361) point out that an assumption is made in scale anchoring, reminiscent of Guttman's scaling principles, that "...examinees know and are able to do all of the skills in the statements at or below a given score level."

As alluded to earlier, Ebel (1962) performed a similar task to what we currently refer to as scale anchoring, with elements of item mapping. Ebel's work focused on the lack of interpretability of test scores if they are not linked to content, and implored his contemporaries to move past normative score reporting. The example data used in the article consisted of 50 SAT® mathematics items, within ten content categories. Ebel (1962, p. 18) proposed "scale books", which he created by sampling the most discriminating item from each content category, resulting in a scale book of ten items. Ebel then divided the examinee pool into six, evenly spaced score categories, and looked at the most common scores on the ten item subset within each score group. The ten item

subset could then be shared with test users, along with the information for likely scores on the sample given overall ability, to add meaning and interpretability to the scaled scores without compromising the entire 50 item assessment. Table 2 from Ebel's work is shown below, displaying the frequency distribution for scores on the 10-item scale book by standard score level/category.

Frequency Distributions of Scores by Standard Score Level on the Ten Representative Items

Score on 10 Items	Standard Score Level					
	750	650	550	450	350	250
10	40	13	1			
9	41	30				
8	16	33	9			
7	3	15	28			
6		8	29	5	1	
5		1	19	15	1	
4			13	24	4	
3			1	28	15	
2				20	30	1
1				8	32	27
0					17	72

Figure 3. Table 2 from Ebel (1962, p. 20)

Standard Setting

Standard setting is the third method described by Kolen and Brennan (2004) regarding the incorporation of content information into the scale score. In standard setting, SMEs or other qualified experts review statements about examinee KSAs, and attempt to locate the point(s) on the score scale that differentiates between the examinees who possess/exhibit the specified KSAs and those who do not possess/exhibit the

specified KSAs. Depending on the purpose of the test, this may be a single scale location, such as instances where an assessment is used to ultimately group examinees into “pass” or “fail” categories, or multiple locations, such as educational achievement tests. The exam program used in this study is a large-scale, medical certification examination, and thereby uses a single cut-score to discern individuals demonstrating the qualities necessary to practice in the field versus individuals who have not demonstrated the necessary qualities.

Many methods have been developed in the interest of setting passing standards and dividing the score scale. The methods can be broadly categorized based on whether SMEs are asked to use test information/data or examinee information/data in recommending their cut-scores between qualitatively distinct performance categories (Jaeger, 1989). One component that nearly all contemporary standard setting methods share is that of human judgment (i.e., regardless of whether the unit of judgment is a test item or a portfolio of student work, SMEs are asked to rely on their expert judgment when setting passing standards). The two contemporary standard setting methods that were used in this study are described next.

The Angoff Method. The initial, rather scant reference to the Angoff method appeared in a 1971 book chapter (Angoff, 1971). Since then, the method has undergone refinements and various “modifications”. Today, the “modified” Angoff method is arguably the most popular standing setting method in operational use (Plake & Cizek, 2012). As Plake and Cizek (2012) describe, it is interesting to note that the method did not come to fruition as the primary subject of a peer-reviewed article or book chapter.

Instead, Angoff put forth the method in a couple of paragraphs and a footnote in his chapter on scaling, norming, and equating for the second edition of *Educational Measurement* (Thorndike, 1971). The basic or unmodified version of the Angoff was the subject of two paragraphs, and what has become the basis for the modified Angoff was espoused in the footnote.

In his 1971 chapter, Angoff describes a process in which panelists are asked to envision a minimally competent examinee for an ascribed ability level, then progress through the items of a test form, deciding whether or not a minimally competent examinee would respond correctly to each item. The items a minimally competent examinee is expected to answer correctly can then be summed to compute the cut-score. Angoff described a variation of this method in his footnote, in which panelists would estimate the proportion of minimally competent examinees to respond correctly to an item, rather than making a holistic yes/no distinction. The expected proportions could then be summed to arrive at the cut-score. What has become known as the *modified Angoff* is the most popular variation of the method used today, relying on panelists to estimate the proportion of minimally competent examinees to correctly answer an item, and doing so over multiple iterations usually in the form of rounds (Plake & Cizek, 2012). Using multiple rounds allows for panelists to refine their ratings, see how their projected cut-scores compare to those of their peers, and receive other feedback (for example, impact data showing percent passing or the percentage of examinees who would fall into each performance category based on their ratings).

Plake and Cizek (2012) point out three important features of the Angoff method that have had a substantial and lasting impact on standard setting. First, Angoff's notion of a minimally competent examinee is a paramount feature in many standard setting methods, so much so that time in the introductory and training phases of standard setting procedures is devoted specifically to developing a conceptualization of the minimally competent examinee. Second, the method relies on SMEs examining test items to make judgments, a hallmark of test-centered standard setting approaches. Lastly, the Angoff method can be applied to tests with multiple cut-scores and instances where multiple-choice items, constructed-response items, or a mixture of item types are used.

To guide the judgmental process and help define the minimally competent examinee (or minimally competent examinee population), panelists can rely on existing PLDs, proficiency claims, or expectations; or these can be developed at the beginning of the standard setting workshop (Plake & Cizek, 2012). If PLDs/claims/expectations are established prior to the standard setting, it may still be necessary to discuss the descriptors in light of the borderline or minimally competent examinee. Adjusting the descriptors for the specific population may also be necessary, as the descriptors are often oriented to the average examinee of a given level (note, level here referring to a performance level, such as "basic" in educational testing or "competent to practice" in certification). Panelist training is a key feature in establishing the validity of standard setting, and will be discussed momentarily.

Despite its overwhelming popularity, the Angoff method is not without its critics. One of the main criticisms relates to panelists' difficulty in the judgmental process of

estimating item difficulties for minimally competent examinees (Plake & Cizek, 2012; Shepard 1995). In a study published in 1998, Impara and Plake used a sample of 26 teachers to investigate SMEs' ability to estimate item difficulty for a population of examinees at large and for a borderline group. The teachers participating in the study were highly familiar with the test content, as well as with the ability of their students. It was found that, while the teachers did a decent job of separating students into the two ability groups, the teachers were not found to produce accurate predictions of the item difficulties for the borderline group of students. Bejar (1983) reported similar results as he investigated whether SMEs could estimate the difficulty and discrimination of items on a written English assessment, and potentially identify the factors that drove the item statistics. The study found that the SMEs did not approach high accuracy in regard to the item statistics, and were unable to isolate the factors that were driving the statistics. However, Bejar (1983) argued that further examination of the items from linguistic and psycholinguistic perspectives to derive the linguistic features present in the items, and coupling this information with SME ratings, could prove fruitful in predicting item statistics.

Reckase and Bay (1999) raised a concern with the yes/no nature of the basic, or unmodified, Angoff judgmental process relating to the range restriction that can occur depending on the selection of items used. Plake and Cizek (2012) used two samples of items with p-values of .7 and .3, respectively, to illustrate this point. If a selection of items is used that is tightly clustered around a p-value of .7 (percent correct of 70%), then it is likely that the panelists will select "yes" for all items resulting in a standard that

corresponds to a perfect score. Similarly, if a selection of items is used that is tightly clustered around a p-value of .3 (percent correct of 30%) , then panelists will likely select “no” for all items and result in a standard of zero correct items. Plake and Cizek (2012) note that range restriction is clearly an issue for the yes/no method; however, limited research has been conducted on the matter.

Impara and Plake (1997) investigated the comparability of cut-scores set using the unmodified and modified Angoff approaches, and found the approaches to yield comparable cut scores. Relating back to the issue of estimating item difficulty, Impara and Plake (1997) also found that panelists were more comfortable with the yes/no procedure than with the modified Angoff approach of estimating percent correct.

Based on the research presented regarding SMEs lack of accuracy in estimating item difficulty (quantified as percent of examinees responding correctly), general comparability of modified and yes/no results, and SME comfort with the yes/no method, the yes/no variation of the Angoff was selected as one of the contemporary methods in this study against which to compare the proposed method. The title of “unmodified” Angoff may not be appropriate as multiple rounds were used; however, SMEs were asked to make holistic yes/no judgments on the items, rather than predict the percentage of minimally competent candidates responding correctly to each item. The risk of range restriction in the yes/no method noted by Reckase and Bay (1999) was mitigated in the study by selecting items that spanned a wide range of item difficulties.

The Bookmark Method. The second contemporary standard setting method used in this study was the Bookmark method, or Bookmark Standard Setting Procedure

(Lewis et al., 1998; Lewis et al., 1996; Mitzel et al., 2000). The Bookmark method was first developed in 1996 in the wake of an influx of statewide standardized testing that required achievement level standards for end-of-grade and end-of-course K-12 examinations (Lewis, Mitzel, Mercado, & Schulz, 2012). The Bookmark method has steadily grown in popularity since its inception, due at least in part to the adoption of No Child Left Behind (NCLB, 2002), which required the reporting of at least three performance categories (Lewis et al., 2012). While the Angoff method can be used to set cut-scores for multiple performance levels, adding performance levels increases the length and expense of the workshop, and could potentially lead to panelist fatigue (Lewis et al., 2012).

The creators of the Bookmark method attribute the roots of the method to item mapping and item anchoring procedures (Lewis et al., 2012). In the Bookmark method, ordered item booklets (OIBs) are created, containing items as they would be mapped along the proficiency scale. Using item response theory (IRT), a response probability (RP) is selected for mapping the items to the scale. A response probability of 67% is often written as RP67, a response probability of 50% as RP50, and so on. Panelists progress through the OIB, and place their cut-scores, as a bookmark, "...between the two items in the ordered item booklet such that from his or her perspective, those items before the bookmark represent content that borderline examinees at a given performance standard should be likely to know and be able to do" (Hambleton & Pitoniak, 2006, p.443). The scale locations of the bookmarks for each cut are then aggregated across

panelists, and the aggregated locations serve as the cut-score recommendations differentiating the levels of proficiency.

The selection of RP has nontrivial implications in the judgmental process undertaken by panelists. As described by Lewis et al. (2012), the first two outcomes realized by the selection of an RP are that a lower RP will result in a lower cut-score, and depending on the IRT model used by the testing program the ordering of items could change depending on the RP. The first point, lower RPs should result in lower cut-scores, is a non-issue as long as panelists are cognizant of the RP when placing their bookmarks. Research has substantiated the claim that lower RPs result in lower cut-scores recommended by panelists; however, they are not in accordance to a level that make the choice of RP inconsequential (National Academies of Science, 2005; Williams & Schulz, 2005). The second issue raised by Lewis et al. (2012), that of potential changes in item ordering based on the IRT model used, can be of minimal or sizeable impact depending on the characteristics of the items used in the OIB. For example, if a two-parameter IRT model is used then the amount of switching that will occur will depend on the range of discrimination values (with more variability in item discrimination parameters leading to more switching) and the separation of the difficulty parameters between items (with more tightly clustered items being more prone to switching). Similarly, if a three-parameter IRT model is used then, varying pseudo-guessing parameters will lead to items changing OIB location with changes in RP.

As the selected RP can have an impact on the outcome of a standard setting, considerable research has been conducted regarding the appropriate RP to use for

ordering the OIB and relating “mastery” to the panelists (see Huynh, 1998; Lewis et al., 1996; Wang 2003). In the literature, RPs of 50, 67, and 80 are often considered. Lewis et al. (2012, p.233) has “discontinued” the use of RP50 when conducting Bookmark standard settings, as an RP of 50 is, “... poorly aligned with the attribution of skills associated with (a) the term mastery in its use during Bookmark and other standard settings and (b) what students in various performance levels should know and be able to do with respect to the PLDs.” Lewis et al. (2012) states that a case can be made for any value greater than RP60, but states that RP80 generally serves as an upper bound and may be too stringent. Lewis et al. (2012) describes how the RP67 was selected in his and colleagues work on the TerraNova (CTB/McGraw-Hill) assessment. The RP allowed for items to be appropriately spaced to cover all performance levels, was anticipated to be aligned with panelists conceptualization of the proficiencies of examinees, and lastly because two-thirds was an easy value to conceptualize between the values of .6 and .8. Lewis et al. (2012) states the practice of delineating PLDs prior to standard setting has quieted the debate over which RP to use.

A benefit of the Bookmark method, as stated by Lewis et al. (2012), is that the Bookmark method can be used to help panelists understand the score scale as it relates to PLDs and expectations of examinees. Lewis et al. (2012, p. 236) states, “The OIB and associated tasks help panelists understand student achievement as a continuum of increasing knowledge, skills, and abilities in the content area.” This view of an ordered proficiency scale is akin to Guttman’s scaling principles, as well as that of AE.

Based on the previous discussion of the Angoff method, some similarities are apparent. Both methods fall under the test- or task-centered distinction as originally described by Jaeger (1989) and Kane (1994). Thus, despite asking panelists to envision minimally competent examinees within an ability level and make judgments as to their abilities, it is ultimately elements of the assessment that are used to set the cut-scores in these methods. Specifically, individual assessment items are used in both methods to set performance standards. Also, both methods can be used to set multiple cut-scores and can be used with selected-response or constructed-response item types.

Naturally, the methods described thus far diverge in many aspects as well. Compared to the Angoff method, one of the primary selling points highlighted by proponents of the Bookmark method is that the Bookmark should require less time by the SMEs to participate in the judgmental process and less data analysis time. The number of data entries for the Bookmark method is equal to the number of the cuts to be set multiplied by the number of panelists (i.e., cuts x panelists); whereas, in the modified Angoff procedure the number of data entries is equal to the number of cuts to be set multiplied by the number of items multiplied by the number of panelists (i.e., cuts x items x panelists). Less data entry is also an advantage of the Bookmark method in regard to the unmodified Angoff; however, depending on the structure of the data, it is possible for the dataset in an unmodified Angoff to only be as large as the number of items multiplied by the number of panelists, regardless of the number of cuts. In the unmodified Angoff method, if an examinee proficiency category is deemed as being able to respond to an item correctly (given a “yes”), then every performance category higher on the proficiency

continuum should also be able to respond to the item correctly as well (also given a “yes”). For example, if the performance categories in ascending order of proficiency are basic, proficient, and advanced, any item that a basic examinee is expected to answer correctly should also be answered correctly by proficient and advanced examinees.

Peterson, Shulz, and Engelhard (2011) reviewed 27 studies conducted by NAEP over the course of 15 years in an attempt to compare the reliability and validity of the Angoff and Bookmark methods. Three criteria were used in the Peterson et al. (2011) evaluation of the methods- the reliability of the set cut scores, the validity of Bookmark set cut-scores by their comparability to Angoff set cut-scores, and procedural validity evidence as quantified by panelist understanding and confidence in the method.

Ultimately, the study concluded that the Angoff and Bookmark methods were comparable in the three criteria investigated, and recommended that, given their comparability, the Bookmark method be used due to the method generally being shorter and less costly.

As noted earlier, the Bookmark method was selected as the second contemporary method to be used in this study. Two primary reasons attributed to this decision. First, the popularity of the Bookmark method position the method as a solid benchmark against which to evaluate the proposed method. Second, the proposed method diverges from the Bookmark method in regard to the unit of judgment; however, the proposed method will share many similarities with the Bookmark method, due in large part to the Bookmark method’s focus on the score scale being an ordered progression of knowledge and skills.

In accordance with popular implementations of the Bookmark method and with the literature provided, an RP67 was selected for use in this study.

Principled Assessment Design

Until now, this chapter has focused on methods for attributing meaning to a scale after the assessment has been developed, and in most cases after examinee response data has been collected. We will now transition to focus on Ebel's (1962) proposition, that of building meaning into the scale during construction.

Discussed previously, the most widely used approach to test development and design revolves around the content blueprint. The content blueprint, as articulated by SMEs and perhaps job/practice analyses, gives a broad overview of the construct of interest. Leighton and Gierl (2007) discuss this approach under the label of a "test specifications" framework, with the test content being represented as a matrix of content areas by cognitive levels. One benefit of this approach acknowledged by Leighton Gierl (2007) and Luecht (2013a) is its ease and simplicity. Content experts and testing organization staff members can create and approve such a model in a relatively straightforward process; however, the simplicity of the model has ramifications in terms of the quality of measurement and inferences that can be made from assessments constructed under such a framework. Leighton and Gierl (2007) point out a limitation of such a model, in that the model is limited to supporting the rank-ordering of examinees and loose inferences as to the skill mastery of the examinee. Similarly, in describing the content blueprint's approach reliance on content validity, Luecht (2013a, p. 5) states,

“Content validity does not provide any direct evidence that aids in the interpretation of scores or inferences drawn from observable performance on a particular form of the test.”

Diverging from the almost exclusive focus on content representation, principled assessment frameworks focus on the properties of measurement scales, the KSAs associated with the different levels of the scale, and how such KSAs manifest and can be measured and demonstrated in reality. Introduced in the first chapter, one of the most well-known principled assessment frameworks is evidence-centered design (ECD; Mislevy, 2006; Mislevy, Steinberg, and Almond, 2003; Mislevy & Riconscente, 2006).

As described in Mislevy et al. (2003), ECD begins with collecting and organizing information about the domain of interest. In the second, domain modeling stage, domain-related information is then organized into *paradigms*. Mislevy et al. (2003) described three types of paradigms- proficiency paradigms, evidence paradigms, and task paradigms. The three types of paradigms are defined below, as they are defined in Mislevy et al. (2003, p. 7):

Proficiency paradigms: structures that organize potential claims about students and aspects of proficiency they reflect.

Evidence paradigms: the kinds of things students might say or do that would constitute evidence about these proficiencies.

Task paradigms: the kinds of situations that might make it possible to obtain this evidence.

The focus of these paradigms is in delineating how KSAs and behaviors manifest across the proficiency continuum, what behaviors will allow examinees to demonstrate such KSAs, and lastly what tasks can be configured that will allow examinees to demonstrate the KSAs if they possess them.

The ECD framework is grounded in evidentiary reasoning (Mislevy et al., 2003), and flows from Messick's (1994) work on evidence and inference. In discussing a construct-centered approach to performance assessment, Messick (1994, p. 16) states, "A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise value by society." The corollary of such considerations manifests as the proficiency paradigms of ECD. Similarly, the ideas of evidence and task paradigms are referenced in Messick's work (1994, p. 16), "Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit behaviors?"

While ECD involves many other aspects of assessment production and delivery, the principles of delineating how proficiency claims are organized along a scale continuum, what manifestations of these claims would like in terms of examinee behavior, and what tasks can be developed that will allow examinee for the evidence of such behavior to be collected are central to the discussion at hand. Luecht's (2006) AE follows a similar path and logic to that of ECD and Messick's (1994) discussions on evidence. As stated previously, AE can be a means by which to implement ECD (Luecht, 2013a), and provides structures for how to go about depicting the construct, ordering of proficiency claims along the construct's scale continuum, and in-depth task specification.

Designing a scale and subsequent assessments from an AE perspective begins with construct mapping. Citing Wilson's (2005) work, Luecht (2013a, p. 8) describes construct mapping as, "... a fundamental design process that lays out the *ordered*

proficiency-related claims that we wish to make about examinees' knowledge and skills at incrementally higher levels of one or more construct-based scales.” Luecht (2013a) emphasizes the ordering of the scale and the importance of performing construct mapping prior to item and task development if possible. Conducting construct mapping first and creating evidence and task models that flow from the delineation of the construct helps to support valid interpretations of scale levels and scores. Luecht (2013a) describes that without such direction it is possible for item and task development to go off track. Such a derailment likely results in wastefulness in the form of tasks and items that do not appropriately align to the scale or support the inferences being made regarding scale values.

Connecting the construct map to the task models are the evidence models (Luecht, 2013a). Evidence models act to supplement the proficiency claims of the construct map by demonstrating what mastery would look like. Similar to the evidence paradigms of ECD, evidence models help to answer the question of, “What products or behaviors do we need to see to know an examinee has mastered the KSAs requisite at this level of proficiency?” Luecht (2013a) describes gathering tangible documents and exemplars of examinee work to guide this process.

As stated, the evidence models connect the construct maps to the task models. As defined by Luecht (2013a, p. 10):

A task model is a cognitively oriented specification for an entire family of items. Each task model integrates declarative knowledge components, relationships among those components, and cognitive skills, as well as relevant content, contexts, and auxiliary features that affect the cognitive complexity of the task.

In specifically stating the components of a task that drive the difficulty of that task, the idea is that task models can be located on the scale based on the cognitive complexity that results from the combination of the components. Furthermore, by creating tasks and items from task models rather than generic content and statistical targets, the evidence demonstrated by an examinee is more explicitly linked to the construct of interest. The visual representation for how task models align to the underlying scale is referred to as the task model map (Luecht, 2013a). While all the potential task models can be used to populate a task model map, different combinations of task models can be assembled to create task model maps for specific purposes. For example, a unimodal task model map would be preferred for an assessment with one cut-score, but a bimodal task model map would be preferred for an assessment with two cut-scores.

The Link between Principled Assessment and Standard Setting

While AE offers many advantages to a testing program, the benefit central to the current discussion is in the attribution of meaning to the scale via explicit design. In this way, AE and standard setting share the purpose of adding interpretability to scale scores. Despite this shared purpose, principled assessment structures have rarely been incorporated into the broader process of standard setting (see Plake, Huff, & Reshetar, 2010, for a study where ECD was used to help develop PLDs), and have not been incorporated into the judgmental process for SME panelists.

In the focus article of a 2013 special issue of *Measurement: Interdisciplinary Research and Perspectives*, Wyse discussed the incorporation of construct maps into standard setting. While Wyse's (2013) definition of construct maps expanded beyond

what we have discussed within an AE framework, Wyse pointed out a key challenge in standard setting, “The challenge in standard setting is often one of seeing the relationships between the PLDs; the scale; and the item, examinee, and rater data on that scale because explicit relationships between concepts are often not made transparent” (p. 142). By incorporating the explicitly delineated knowledge and cognitive skill components of task models into standard setting, the author hopes to aid panelists in bridging the gap between the demands of a task and their expectations or proficiency claims regarding a minimally competent candidate.

The Validity of Standard Setting

In this section, literature will be reviewed that influenced how the standard setting workshops in this study were conducted. The framework by which the validity of standard setting workshops is evaluated will be used to direct the discussion. Framing the discussion in such a way will allow for explanation as to why the standard setting workshops in this study were conducted in the manner that they were, and provide an opportunity to highlight where the proposed method seeks to bolster the validity of the outcomes of standard setting.

Standard setting does not only have implications for the validity of test scores as defined by Messick (1989), but frameworks have been developed for examining the validity of the standard setting process itself (see Kane, 1994; Norcini & Shea, 1997; Kane 1999; Pant et al., 2009). The frameworks for collecting validity evidence have generally broken the evidence into the categories of procedural, internal, and external, with Pant et al. (2009) further dividing the external category into external and

consequential. While the author acknowledges the importance of the consequential validity element of standard setting, it will not be addressed in this study, and thereby will be omitted from discussion here. The procedural, internal, and external elements had the most direct implications for how the standard setting workshops were implemented in this study. Each element will be summarized, and areas where debate exists among best practices will be discussed in further detail (for example, the selection of panelists within the procedural validity element), in order to explain the reasoning behind design decisions of the current study.

Procedural

As the name would imply, procedural validity evidence is concerned with, “the appropriateness of the procedure used and the quality of the implementation of these procedures” (Kane, 1994, p. 437). The procedural validity evidence for standard setting studies can be further broken down into five key components- the explicitness of the plan and procedures, the practicability of the procedures, the implementation of the procedures, panelist evaluations, and proper documentation (Hambleton & Pitoniak, 2006; Pant et al., 2009). The explicitness of a standard setting refers to how well the processes are articulated prior to the standard setting. The practicability of a standard setting refers to how feasible the methods are given real world constraints, and how communicable the methods and results of a standard setting are to participants and laypeople. The implementation of the procedure refers to how well the a priori plan was followed, and to the extent panelists were appropriately trained, performance levels were defined, and data was systematically collected. Panelist feedback is instrumental in

establishing procedural evidence, in that it allows for the panelists' to share their perspective on the efficacy of the training, the appropriateness of the judgmental process, and the confidence the panelists' had in their judgments. Documentation, the last aspect of procedural evidence, refers to how well the overall practice is documented.

Panelist selection- expertise. Hambleton and Pitoniak (2006) discuss two primary factors in selecting panelists for a standard setting procedure- the qualifications of the panelists and the total number of panelists. For assessments that have a wide range of stakeholders, it can be difficult to know or decide who should be enlisted to conduct the standard setting. The composition of the standard setting panel can have a profound effect on the set cut-scores, as Jaeger, Cole, Irwin, and Protto (1980) found large discrepancies in the percentage of North Carolina students who would pass two assessments depending on whether educators, the general public, or state legislators set the standard.

Hambleton and Pitoniak (2006, p.451) state that a common approach to assembling a standard setting panel is to include, "... all groups that have a legitimate stake in the outcome of an assessment and the decisions that will derive from its use." Many scholars share the opinion of including a broad range of interests in the standard setting panel, including Cizek (1996b), Hambleton (1978), Kane (1994), and Shepard (1980). Specifically, Shepard (1980, p.454) gives the advice, "... ensure that different value positions and areas of expertise are systematically represented when judges are empanelled." In Loomis' (2012) chapter, she describes how NAEP selects standard

setting panels of 70% educators and 30% non-educators to achieve the right mix of stakeholder groups.

The importance of panelist background and expertise in the subject area can be magnified in the test-centered approaches, especially in instances where panelists are asked to estimate examinee performance on a set of items. In his review of panelist expertise literature as it relates to item difficulty estimation in the modified Angoff procedure, Brandon (2004, p.66) found that panelist expertise can play a role in item difficulty estimation, but ultimately concluded, "... expertise can enhance item estimation, but that not all judges need to have high levels of expertise." Brandon (2004) also found that providing panelists with empirical item difficulties and holding discussions between rounds mitigated the differences in item difficulty estimation between panelists of varying expertise. The body of literature appears to suggest that including a range of stakeholders may be advantageous in setting realistic and feasible performance standards. As the cut-scores set during a standard setting usually serve as "recommendations" to policymakers, including administrators, policymakers, and other stakeholders can assuage the disparity that may result between recommended cut-scores and policymakers expectations. Furthermore, Shepard (1980) recommends summarizing the standard setting results by panelist subgroups to demonstrate the variability in the judgmental process.

In reviewing the literature on panelist expertise, it is clear that there are merits to including a mix of content experts and stakeholders in a standard setting workshop. However, the material on the examination used in this study requires not only medical

knowledge, but is designed for examinees who have completed medical school and at least three years of a fellowship program specific to the field of study. Due to the extremely specialized content of the examination used in this study, it was decided to only include experts in the field when recruiting for the standard setting workshop. As pointed out in Hambleton and Pitoniak (2006), content knowledge is the most important criteria in selecting panelists.

Panelist selection- number of panelists. Hambleton and Pitoniak (2006) equate trying to determine the appropriate number of panelists for a standard setting to the common research question of appropriate/required sample size. As standard setting procedures fall victim to many practical concerns in regard to resources, achieving a large number of panelists can be difficult. Cizek (1996a) and Hambleton and Pitoniak (2006) state the best approach may be to empanel as many panelists as resources will allow; however, such a number could still be determined to be insufficient, or could result in more precision than the process requires. Kane (1994, p.441) recommends, "... the number of judges should be large enough to achieve an acceptably small standard error of measurement for the resulting passing score." However, as Hambleton and Pitoniak (2006) note, the standard error of the cut-scores can depend on many factors and be difficult to anticipate. Varying numbers of panelists have been recommended in the literature. Norcini and Shea (1997) discuss stable standard results with five to 10 panelists. Zieky and Livingston (1977), as cited in Brandon (2004), suggest a minimum of five panelists but no more than 30. In their work on legal defensibility of high-stakes

testing, Mehrens and Popham (1992) suggest empanelling 20 to 25 panelists in an effort to keep the standard error of the mean low.

Generalizability theory has also been used to help determine the sample size of future standard setting procedures, specifically using the dependability coefficient (ϕ). Norcini, Lipner, Langdon, and Strecker (1987) found dependability coefficients ranging from .79 to .94 with as few as five raters using variations of the Angoff method. In a study that compared the Angoff method to two other procedures for a teacher certification exam, Cross, Impara, Frary, and Jaeger (1984) found coefficients ranging from .53 to .82 depending on the round for the Angoff method (six rounds in total, two subjects with three rounds each). These coefficients were based on five participants, and Raymond and Reid (2001) estimate the most dependable of the conditions would be expected to have a coefficient of .88 if the number of participants were increased to 10, and .92 if increased to 15.

In conclusion, it is generally recommended that as many panelists as resources will allow should be empanelled, within reason. Studies have shown that acceptable levels of reliability or dependability can be achieved with as few as five participants; however, samples of closer to 10 or more are generally desirable. Based on this information and given the operational constraints of organizing three unique groups of panelists for the current study, the current study invited 10 SMEs to participate in each of the three standard setting method groups.

Training panelists. Citing Mills (1995) and Raymond and Reid (2001), Hableton and Pitoniak (2006) outline four components that should be addressed in

training participants for standard setting. First, the general process the panelists will be undertaking should be reviewed with the panelists, along with a schedule so panelists are aware of the time constraints on given activities. Second, the purpose, use, consequences, and other contextual information regarding the standard setting and larger assessment should be discussed with the panelists. Along the lines of consequence, Raymond and Reid (2011, p. 142) state, “The facilitator should encourage participants to discuss the consequences of setting a standard at a particular score level, as well as the consequences of failing or passing the examination.” Raymond and Reid (2011) also recommend describing the difference between norm-referenced and criterion-referenced assessments, as some panelists may only be familiar with the former. Third, the panelists need to develop a conception of the reference group (Hambleton & Pitoniak, 2006). Regardless of when the PLDs of examinees (or expectations of candidates in certification/licensure) are developed, it is important for the panelists to have a clear idea of the KSAs of the target population, with specific attention to the examinees falling into the minimally competent region of performance levels. Fourth, and lastly, panelists should be trained in the specific judgmental process they will be using (Hamblteon & Pitoniak, 2006). In addition to training, it can be useful to participate in an abbreviated practice round resembling the actual judgmental process the panelists will be using. In accordance with these best-practices, all four components were incorporated into each standard setting session used in the current study, and will be addressed in more detail in the methods section.

Providing Feedback to Panelists. Providing feedback to panelists is the last topic that will be discussed within the procedural validity element of standard setting. Providing feedback has become a common practice in standard setting procedures (Hambleton & Pitoniak, 2006). As most standard setting methods have either been designed or adapted to include multiple rounds, the gaps between rounds provide the opportunity to share information with panelists and allow for discussion. Common forms of feedback given to panelists between rounds include how their ratings or placement of cut-scores relate to their peers, empirical item or task difficulties if they have not already been provided, normative data in the form of impact data, and active feedback in the form of group discussion (Hambleton & Pitoniak, 2006).

Hambleton and Pitoniak (2006) point out that the merits of providing feedback to panelists have been debated. Policymakers may be opposed to providing normative or feedback data to panelists so as to not contaminate a process geared toward criterion-referencing and KSAs with norm-referenced information. On the whole, it seems that panelists value feedback, “Panelists often report that they appreciate the opportunity to view feedback data and to discuss their ratings with their colleagues, and find these activities valuable” (Hambleton & Pitoniak, 2006, p. 456).

As providing feedback to panelists has become common practice in standard setting workshops, feedback and discussion were incorporated into the standard setting workshops of this study. Feedback was provided in the form of how panelist’s cut-scores compared to one another’s, impact data showing the anticipated pass/fail rates if the recommended cut-score was adopted, and group discussion was held. The author

acknowledges that incorporating impact data has the potential to inject norm-referenced thinking into a primarily criterion-referenced process. That said, impact data was ultimately included in an effort to make the standard setting workshops as realistic as possible, and mirror how the workshops would be conducted outside of a research setting.

Internal

Generally speaking, the internal sources of validity evidence refer to the reliability of a standard setting procedure. Hambleton and Pitoniak (2006) identify four key components for compiling internal validity evidence- the consistency within the standard setting method, intrapanelist consistency, interpanelist consistency, and consistency across components of the assessment and its items (i.e., consistency across content subdomains, item types, etc.). Hambleton and Pitoniak (2006) cite Kane (2001) as describing two methods for evaluating consistency within a standard setting method. First, the standard error of the mean could be estimated using Brennan's (2002) formula when the same method was employed on two separate occasions:

$$\sigma_{\bar{x}} = \frac{|\bar{X}_1 - \bar{X}_2|}{2},$$

where \bar{X}_1 is the standard from the first implementation of the method and \bar{X}_2 is the standard from the second implementation of the method. Second, generalizability theory has been recommended due to the practical difficulties of convening multiple standard setting sessions of the same method.

Intrapanelist consistency, or within panelist consistency, can be looked at in two ways, as described by Berk (1996). The first case is within panelists across the rounds of a standard setting. Interpreting consistency in such a context can be confusing, as high or low variability (corresponding to low or high consistency) may be ideal under different circumstances. For example, a high level of variability within a panelist between rounds could mean that the panelist is using the feedback and information provided between rounds to refine their ratings. In such a case, high variability, and thereby low consistency, would be a positive for the validity of the process. Conversely, if a panelist is confident in their ratings and the feedback/information provided between rounds is serving to confirm their ratings, the low variability and a high coefficient of consistency would be obtained. For these reasons, determining why the variability within a panelist between rounds was either high or low is crucial if anything above descriptive statements are to be made regarding panelist consistency between rounds. The second case in examining intrapanelist consistency requires using auxiliary information to gauge a panelist's consistency within a round. Unfortunately, such a check may or may not be doable depending on the standard setting method. For example, such an investigation is applicable in the modified Angoff method, where the variability in a panelist's recommended cut-score can be addressed by looking at the scale score corresponding to each proportion correct estimate for each item; whereas, in the Bookmark method panelists only make one decision for each cut-score, not allowing for variability within panelists within round.

Berk (1996, p.229) refers to interpanelist consistency as the, “degree of homogeneity or internal consistency of the final decisions by the judges.... It is essentially an index of convergence.” Berk (1996) advocates the use of generalizability theory for computing the level of interpanelist consistency. Berk (1996, p.229) gives example reasons for why interpanelist consistency may be low, specifically, “...ambiguity in definitions of achievement levels, format of items or exercises, competence of judges, background characteristics of judges”. Hambleton and Pitoniak (2006) point out simple standard deviations can also be used to characterize the variability between cut-scores set by the panelists.

For the current study, interpanelist was assessed by looking at the variability of the final cut-score recommendations for each group. Specifically, simple standard deviations were calculated in accordance with Hambleton and Pitoniak (2006). Visual representations of the data were also be used to depict the distribution and variability of the cut-scores, helping to identify the impact of extreme panelists.

External

While the previous two elements of validity evidence dealt with the process and internal outcomes of a standard setting, external validity evidence for a standard setting is focused on how the resulting performance standards relate to other measures. Pant et al. (2009) partitioned external validity evidence and consequential validity evidence. The sources of external validity evidence discussed in Kane (1994) and Hambleton and Pitoniak (2006) that remained in the external category will be discussed here.

First, external evidence of validity can be collected by comparing the resulting standards of a standard setting method to the resulting standards of another standard setting method. In the third edition of *Educational Measurement*, Jaeger (1989) summarized multiple studies that compared standards across methods. Jaeger (1989) concluded that standards are often found to be inconsistent across methods, and often methods are not implemented identically across settings. Kane (2001) pointed out that it is not surprising for different methods to yield different standards as the judgmental process varies across methods, as well as the information and data used in the judgmental process. In response to this information, Hambleton and Pitoniak (2006, p.461) state:

If the discrepancy between the two performance standards is too large, however, it may call into question the validity of the results from both standard setting processes. In contrast, if different standard-setting methods produce identical or very close performance standards, our level of confidence in both sets of results is increased.

Regardless of whether multiple standard setting methods produce the same results, Zieky (2001) pointed out that comparing standards set by multiple methods is still useful. Wyse's (2013) push for incorporating construct maps, specifically empirical construct maps that can include cut-score recommendations from multiple standard setting methods, coincides with the view that different methods can provide different information, and prove useful to policy-makers in setting standards.

The second source of external validity evidence is in the form of comparing set performance standards to other sources of information. External validity evidence of this nature generally takes on the feel of traditional criterion-related validity evidence, in

which scores or categorization based on one assessment is compared to that of a criterion or synonymous measure. Kane (2001) discusses the most direct way to establish such validity evidence, in which examinees would take an assessment then be evaluated while performing the activity for which the assessment is attempting to predict. However, Kane (2001) recognizes issues with such a design, primarily that the validity of the performance activity would also need to be established, and in some instances it may not be possible to have examinees who did not pass the exam perform the activity. The overall difficulty of collecting criterion-related, external evidence has led to a dearth of this form of evidence in most instances of standard setting (Kane, 1994). While this will not be attainable in the current study, the first piece of external validity evidence (consistency between standard setting methods) will be evaluated in this study by comparing the results of the three standard setting workshops.

Conclusion

In this chapter, relevant literature pertaining to the meaning of measurement scales, standard setting, and principled assessment frameworks was examined. While principled assessment frameworks and standard setting may seem like unrelated topics, the two share a common goal in making scale scores meaningful and useful. The next chapter describes the study that was conducted to evaluate the effect of incorporating AE task models into standard setting.

CHAPTER III

METHODS

This chapter outlines a method of standard setting that uses task models as the unit of judgment. To evaluate the merits of this method, it will be compared against the two prominent standard setting methods described so far in this study- the Angoff and Bookmark methods.

Task Model-Based Standard Setting

The standard setting method proposed in this study uses the cognitively based AE structures of task models to set cut-scores. For ease of reference, the proposed method will be referred to as Task Model-Based Standard Setting (TMSS). Using Jaeger's (1989) and Kane's (1994) classification schema, the method would best be classified as a task- or test-centered method. As the proposed method will also share some similarities with what Brennan (1995) describes as scale-centered anchoring approaches, the method may fall into a new, bridged categorization of "scale-centered standard setting". Brennan (1995, p. 271) stated scale-centered approaches, "...start with a scale and then produces a definition." One of the goals of the proposed method is to move away from singular test form or item instantiations of the construct, and focus on the skills and knowledge that constitute the construct.

As the name suggests, task models are the key components the standard setting panelists will use in their judgmental process. Analogous to how individual items are used in the Angoff and Bookmark methods, individual task models served as the basic unit panelists will evaluate in the judgmental process. TMSS shares many similarities with the Bookmark method in terms of materials and the judgmental process (although the unit of judgment shifts from the item to the task model). Task models were organized in an ordered booklet based on the RP67 of each task model's corresponding item. Similar to the Bookmark method, panelists will progress through the booklet and place their cuts where a switch occurs such that an examinee of the current performance level would no longer be able to perform the task with consistency (quantified as the RP, or roughly two-thirds of the time for an RP67). As with Guttman-like scaling principles and conjunctive scale properties, the higher ability group is assumed to have the requisite KSAs to perform all of the task models prior to the cut.

Task models were represented in the TMSS ordered booklet in a narrative format rather than a task model grammar (see Luecht, 2013a) for the ease of comprehension by panelists, as they may not have been familiar with the format of a task model grammar prior to participating in the standard setting. Furthermore, an example item for each task model was provided to supply the panelists with a concrete example. Presenting an item along with the task model could lead to panelists ultimately using the item, rather than the task model, as their unit of judgment; however, it was decided that including the items to help panelists comprehend the task models was worth the risk of contamination.

As only a single item is displayed per page in an OIB for the Bookmark method, only a single task model was displayed per page in the TMSS ordered booklet. However, as described in the previous paragraph, an example item accompanied each task model. The ordered booklets for the Bookmark and TMSS groups in this study will be discussed in further detail momentarily. Aspects of implementation, such as the training of panelists and the feedback/discussion between rounds, were consistent with standard practice for standard setting workshops, and will be discussed as the implementation of each method is reviewed.

The Study

Three independent standard setting panels were convened to compare the TMSS with the Angoff and Bookmark methods. Panelists were randomly assigned to each group initially; however, three panelists had to switch groups based on scheduling conflicts. It should be noted that it was not logistically feasible to have a single panel engage in more than standard setting method, even though that type of design may have been somewhat more powerful in a statistical senses. However, each group reviewed the same set of items. The research questions addressed by this study are listed below. A brief description of the specific analysis allowing for each research question to be evaluated is listed below each question. Following the research questions, the study is outlined in more detail.

Research Questions

- 1) How do the cut-scores produced by the TMSS, Angoff, and Bookmark methods compare in regard to rigor?

Analyses | The point estimates of the final cut-scores will assess the comparative rigor of the cut-scores.

- 2) How do the cut-scores produced by the TMSS, Angoff, and Bookmark methods compare in regard to interpanelist consistency?

Analyses | In accordance with Hambleton & Pitoniak (2006), the standard deviations of the final cut-scores will be used to assess interpanelist consistency.

- 3) How does the TMSS compare to the Angoff and Bookmark methods in regard to panelists' confidence in set cut-scores, ability to differentiate between different levels of the scale, and ability to describe the competencies of passing candidates?

Analyses | Quantitative analysis of relevant evaluation survey questions completed by the participants, compared across standard setting methods. Also, the final question of the evaluation survey asked panelists to list statements that can be made regarding the competency of passing candidates, given their cut-score and the information they reviewed. If able, the panelists were asked to list three example statements. The responses were analyzed for whether or not the panelists were able to provide relevant information, and whether the statements spoke to the KSAs of candidates.

The Assessment and Items Selected for this Study

The assessment used in this study is the initial certification exam for a large-scale, medical certification body. Candidates sitting for the exam must meet licensure

requirements in the field, and have completed requisite training at an accredited institution. The same subset of 100 items from one form of the 2014 certification exam was used for each of the three standard setting methods. The subset of items was selected to meet the content blueprint of the exam, span the range of difficulty with minimal score gaps, and only items with point-biserials greater than .10 were considered for selection. Figure 4, below, shows the Rasch item difficulties plotted against their position in the OIB. The subset of items was found to be more difficult, on average, than the operational forms administered in 2014. Figure 5, below, shows the distribution of Rasch item difficulty measures.

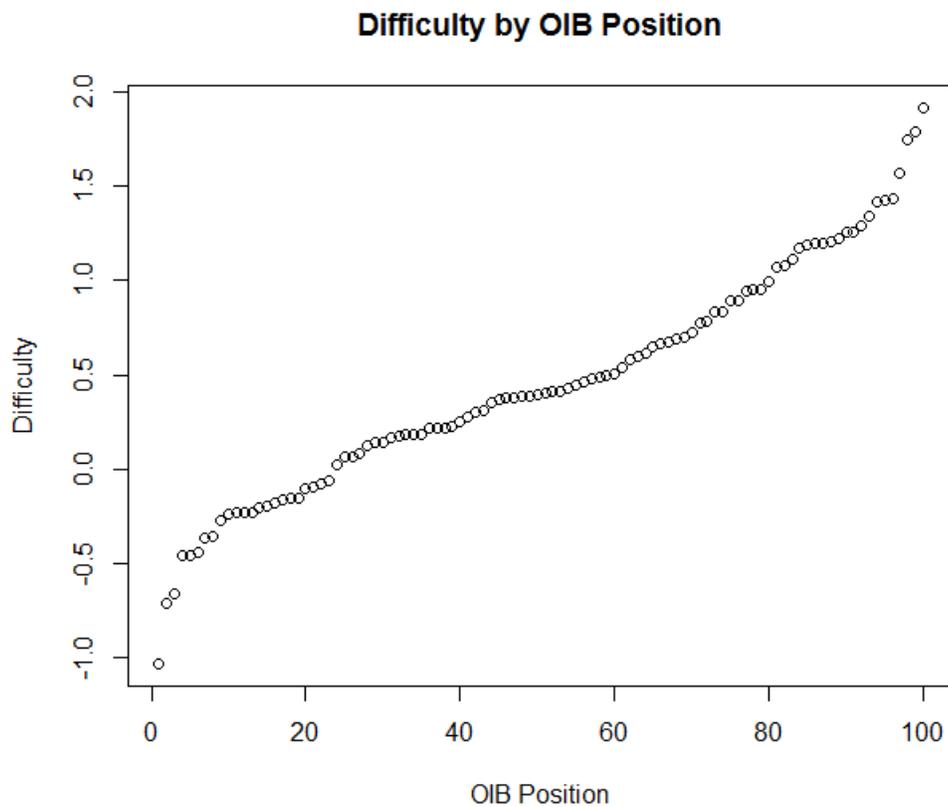


Figure 4. Item Difficulty Parameter by Location in OIB

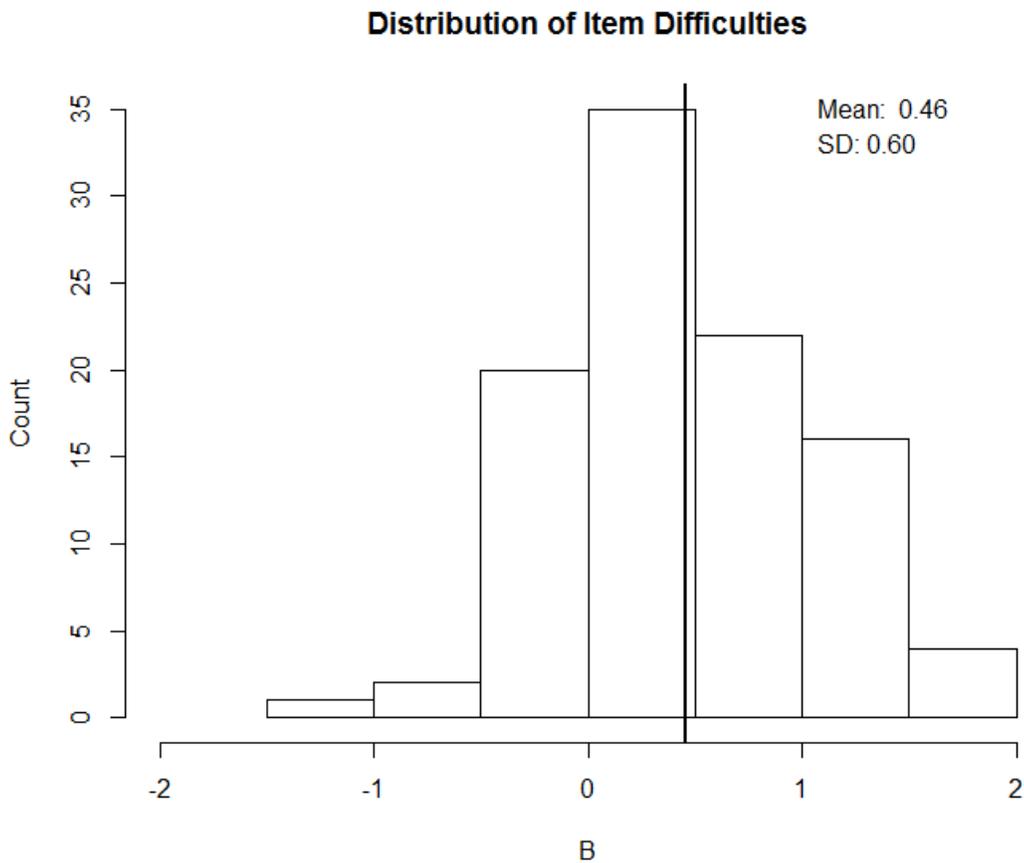


Figure 5. Distribution of Item Difficulties for Standard Setting Items

Creating Task Models

Although a full discussion of the more comprehensive task model creation process is outside the scope of the current study, the basic process by which the task models used in this study were created is summarized below. The exam program used in this study had not previously used AE or task models, therefore cognitively based task models had to be developed in order to conduct the TMSS.

A team of 17 SMEs was formed in order to create task models that reflected the complexity features of the 100 items selected for use in the standard setting workshops. All of the SMEs participating in the creation of task models were currently certified at the time of the study, had maintained their certification for at least 10 years, were considered experts in the field, and participated in the project as volunteers. All of the SMEs were either current or previous members of one or more of the organization's various committees. The organization used in this study had recently instated term limits on committee membership, and the task modeling project provided an opportunity for members rotating off of committees to continue their service to the field. While the SMEs creating the task models were considered experts in the general field, they also represented a sampling of experts from the various sub-specialties within the field.

Introductory webcast. The task modeling project began with a pre-recorded webcast that the task modeling team were asked to view at their leisure, prior to beginning the next phase of the project. The webcast was approximately 30 minutes in length, and provided background information regarding AE and task models.

Live webinar. A live, 2-hour webinar was scheduled following the introductory webcast. The focus of the webinar was to address any areas of confusion, allow the task modeling team to ask questions, discuss features of task models, and to create an initial task model grammar. As stated, the task models would not be presented to the standard setting panelists in task model grammar format; however, it was important to define the schema the SMEs would use in reviewing the items and creating task models.

Task model grammar refinement. In the two weeks following the live webinar, the task modeling team members were asked to sample through some of the items that would be task modeled, and evaluate how well the initial task model grammar would work with the items. Email exchanges between the author and the team members served to refine the task model grammar. Ultimately, the task model grammar that was used in evaluating the standard setting items consisted of three main areas: content information/knowledge, cognitive skill or process, and information density.

Item review from a task modeling perspective. After the task model grammar was refined to a point where the group was comfortable reviewing items under the decided framework, each member of the group was assigned a selection of items. For each item, a team member was asked a set of questions. In regard to content, the questions focused on what information was provided in the stem that was relevant to responding to the item, and what information the candidate needed to recall from his/her medical training in responding to the item. The reviewer was also asked if they felt the content information would be categorized as basic medical knowledge (medical knowledge that should be common to all medical professionals) or distinguished medical knowledge (medical knowledge that would only be expected of medical professionals with specialized training). In regard to the cognitive process required by the item, the reviewer was asked to describe the cognitive process the candidate goes through in responding to the item, and how they would categorize the cognitive process based on an agreed upon, 3-tier system. Lastly, the reviewer was asked to rate the information

density of the item, which was operationalized as low (five or less pieces of information the candidate needed to organize) or high (more than five pieces of information).

Assembling the task models. Once the task modeling team had completed their reviews, the information was organized into a task model statement in the form of a narrative. Due to test security concerns, the task models cannot be shared at this time; however, the general format was as follows:

Given [description of patient; such as male of age X] **presenting with** [relevant content information in the stem; such as a list of symptoms, lab values, etc.],
recall [relevant content knowledge from training; such as relationship between a set of symptoms] **and** [final cognitive process action; such as recognize likely diagnosis of Y, or recognize likely diagnosis of Y and plan appropriate management of patient].

Evaluation of task models. As a final evaluation step, each task model statement and other information from the initial review were evaluated by two other members of the task modeling team. In total, this meant that each item and task model was seen by a total of three members of the task modeling team. In cases where disagreements existed as to what to include in the task model, the content category, or the cognitive category, the opinion of the majority (two out of the three team members) was used as the final decision. This concluded the creation of the task models for use in the TMSS.

Participants

To conduct this study, three committees of 10 participants each were scheduled to participate in the study. Due to unforeseen travel and personal circumstances, three of the 30 panelists were unable to participate on the day of the study. As a result, ten panelists participated in the Bookmark method (seven female, three male), nine panelists participated in the TMSS (three female, six male), and eight panelists participated in the Angoff method (five female, three male).

All of the panelists recruited to participate in the study were considered to be experts in the field. At the time of the study, all of the panelists were actively maintaining their certification in the field. Twenty-six (26) of the panelists recruited to participate were current members sitting on one of the organization's committees, with the remaining four panelists having previously worked with the organization on various projects and committees. As such, all of the panelists were familiar with the testing program/content. None of the panelists involved in the study had conducted their specific standard setting method previously (e.g., while members of the Bookmark group may have participated in standard setting before, none had participated in a Bookmark standard setting). Lastly, while the panelists were all considered experts in the general field assessed by the examination used in this study, the panelists also represented a broad sample of sub-specialties within the broader field (e.g., endocrinology, gastroenterology, etc.).

Schedule

The general agenda is shown in Table 2, below. Each criterion-referenced standard setting method consisted of an introduction and training session, two rounds of standard setting with a break for discussion between rounds, and concluded with an evaluation survey. A Hofstee exercise (1983) was also conducted with each group, after the group completed their evaluation survey for their specific method. Although the Hofstee method was not included as one of the methods of interest for this study, the results will be discussed briefly in the next chapter.

As the Hofstee was the only common method to all three groups, the results of the Hofstee exercise may help to speak to the comparability of the groups. The Hofstee is known as a compromise method (Hambleton & Pitoniak, 2006), meaning it attempts to incorporate both criterion-referenced and norm-referenced information from each panelist to arrive at a cut-score. After reviewing a selection of exam items, panelists are asked to provide their minimum and maximum acceptable passing standard for the criterion-referenced portion. As for the norm-referenced portion, panelists are asked to provide the minimum and maximum acceptable failure rates from a policy perspective. All four values are then averaged across panelists, and used in conjunction with a cumulative density function of examinee test scores to arrive at a recommended cut-score.

While the norm-referenced information is not relevant to the current study, the criterion-referenced estimates provided by the panelists could offer a look into how the panelists viewed the overall difficulty of the selected items, and likewise provide some insight into the reasonableness of their ultimate, criterion-referenced cut-score. For

example, if the TMSS final cut-score fell at a score of 90%, but the Hofstee ratings showed a holistic view of an appropriate cut-score between 70%-80%, this could indicate a disconnect between the group's item-level judgments and their holistic view of the content.

Table 2. Agenda for Standard Setting Groups

Session	Purpose	Time (minutes)
Introduction	Introductions and training session	30
Round 1	First round of cut-scores	90-120
Analysis and Discussion	Share first round cut-scores, impact data, and hold discussion	20-30
Round 2	Second round of cut-scores	30-60
Survey	Final survey closes out study methods	5-15

Introduction. The introduction session followed the recommendations of Raymond and Reid (2001) described in the previous chapter. Specifically, 1) panelists were informed as to the upcoming process and schedule, 2) the purpose, use, and consequences of the standard setting workshop were discussed, along with a discussion regarding criterion-referenced compared to norm-referenced assessment, 3) the KSAs of the reference group (the minimally-competent or borderline candidate) were discussed, and 4) panelists were trained in their specific method, and participated in a three item

practice round before moving forward. The introduction session presentations for all three groups were identical, with the exception of the training on the judgmental task, which was unique to each group.

Round 1. Following the introduction, each group conducted their first round of ratings. For the Bookmark and TMSS groups, this entailed progressing through their item booklets, and noting the page of their cut-score on a ratings sheet. The ratings sheets can be found in Appendix A. The item booklets for the Bookmark and TMSS were identical, with the exception that the TMSS booklet contained a task model on each page, in addition to the operational item. For the Angoff group, Qualtrics survey software was used to capture ratings. Each participant progressed through their item book, and for each item gave a “yes” or “no” response in Qualtrics as to whether the minimally-competent candidate would respond correctly to the item. A screenshot of the Qualtrics survey can be found in Appendix B. The Angoff item book contained multiple items per page in some instances. For all three groups, the key for the selection of items was provided on the first page of the book; however, panelists were asked to only use it if necessary.

Analysis and Discussion. Initial cut-scores were calculated and a group discussion was held for each group following Round 1. For the Bookmark and TMSS methods, cut-scores were calculated by finding the RP67 scale location corresponding to each panelist’s recommendation, then averaging across the RP67 locations. For the Angoff method, an individual participant’s cut-score recommendation was found by summing their “yes” responses, and the initial cut-score for the group was found by

averaging across panelists. After displaying the anticipated pass/fail rate based on their initial cut-score recommendations, group discussion focused on items the panelists wished to discuss. Additionally, the facilitator for each group identified some items for discussion. In the Bookmark and TMSS groups, discussion was directed to the extremes of the recommended cut-scores, as well as areas where panelists felt they went back and forth on their ratings. For the Angoff group, discussion was directed to items where roughly half of the panelists said “yes” a borderline candidate would respond correctly.

Round 2. Following the group discussion, Round 2 of standard setting ratings commenced. Panelists were asked to revisit their ratings, and make any changes they felt appropriate. This concluded the ratings for the three standard setting methods of interest in this study.

Survey. A final evaluation survey was administered at the conclusion of Round 2. The survey for the Bookmark and Angoff methods consisted of eight Likert-type items, followed by a single open-ended question. The Bookmark and Angoff surveys can be found in Appendix C. Statements 1-8 speak to the procedural validity of the standard setting procedure. The open-ended question at the conclusion of the survey was intended to assess how well the panelists were able to characterize the KSAs of passing candidates based on their cut-scores. The survey for the TMSS group can be found in Appendix D. The survey is identical to the other group surveys, but contained three additional pieces of information. Panelists in the TMSS were asked to indicate if they understood task modeling as it was described in the standard setting; if they understood the individual task models used in the standard setting; and then finally if they used the task model, the

item, or a combination of the task model and item when making their final cut-score judgments.

Conclusion

In this chapter the proposed method, TMSS, was outlined. The proposed method shares many similarities with the established methods described, but departs in the sense that the focus of the proposed method is in the task model embodiment of KSAs, rather than items assumed to represent the content of the construct of interest. The study outlined here will allow for the procedural and internal validity to be compared to that of the modified Angoff and Bookmark methods. Panelists' recommended cut-scores will serve as the data in analyzing the first research question, and data gathered from the evaluation surveys of each group will be used to analyze the second and third research questions.

CHAPTER IV

RESULTS

In this chapter, the results of the three standard setting workshops are summarized, with attention to each of the research questions presented in the previous chapter. First, the procedural validity evidence collected in the final evaluation survey will be addressed for each of the standard setting methods. Next, the recommended cut-scores and distributions of cut-score recommendations for each group are presented to address the first research question. Finally, the results of the final evaluation surveys that relate to the second and third research questions are presented.

Procedural Evidence for the Validity of the Workshops

The statements that were specifically used to demonstrate the procedural validity of the three methods will be discussed here, and the statements relevant to the second and third research questions will be discussed later in this chapter. Table 3 contains the frequencies for statements 1-10 of the final evaluation survey, to which participants were asked to respond with Strongly Disagree, Disagree, Agree, or Strongly Agree. Given the structure of the data and the sample size, frequencies are the most appropriate way to analyze the data; however, Figure 6 displays the average response value for each group to each statement. The average values should be interpreted with caution, but can help to identify pervasive differences between groups.

Table 3. Final Survey Response Frequencies

#	Statement	Method	SD	D	A	SA
1	The training for this standard setting method was clear to me.	Angoff	0	0	1	7
		Bookmark	0	1	7	2
		TM	0	0	6	2
2	I understood the goal of this standard setting.	Angoff	0	0	1	7
		Bookmark	0	0	4	6
		TM	0	0	3	6
3	I am confident I applied the standard setting method appropriately.	Angoff	0	0	3	5
		Bookmark	1	1	8	0
		TM	0	0	7	1
4	I am confident my final cut-score appropriately distinguishes between competent candidates and candidates who are not competent to practice.	Angoff	0	1	5	2
		Bookmark	1	2	5	0
		TM	0	3	5	0
5	The discussion between rounds was helpful to me in completing my second round of ratings.	Angoff	0	0	2	6
		Bookmark	0	1	5	4
		TM	0	0	5	3
6	I found the distribution of recommended cut-scores (minimum, maximum, and mean/median) following round one useful.	Angoff	0	0	3	5
		Bookmark	0	1	6	3
		TM	0	0	6	2
7	The information showing the distribution of examination scores was helpful.	Angoff	0	0	3	5
		Bookmark	0	0	5	5
		TM	0	2	3	2
8	I am confident in my ability to distinguish between the knowledge, skills, and abilities of candidates who achieve different scores (e.g., the difference between a candidate who scores 40 versus a candidate who scores 60).	Angoff	0	0	5	3
		Bookmark	0	6	4	0
		TM	0	2	6	1
9	I understand the concept of a task model as it was defined in this standard setting.	TM	0	0	7	2
10	I understood the individual task models that were presented in this standard setting.	TM	0	0	7	2

Shown in Table 3, there is strong procedural evidence for the validity of each of the standard setting methods used in this study. With few exceptions, we can see that panelists understood the training for their standard setting method, the goal of their standard setting, and were confident they applied their method appropriately. Similarly, across the methods nearly all panelists found the group discussion and data shared between rounds (impact data and the distribution of recommended cut-scores) to be helpful in formulating their ratings. All panelists in the TMSS group indicated they understood task models as they were explained in the introduction and the actual task models in the standard setting. Not shown in the table, the TMSS group was also asked if they used the item, task model, or a combination of the item and task model to make their final decision. Of the nine panelists, five (5/9) indicated they ultimately used the item, and four (4/9) indicated they used a combination of the item and the task model.

Figure 6, shown below, displays the average level of agreement for each group, and a few trends are apparent. It is clear that the Angoff group had the highest level of agreement, on average, with the statements in the survey. In contrast and with only one exception, the Bookmark group had the lowest level of agreement, on average, with all statements. The one exception in this case being statement #7 regarding the displaying of impact data during the group discussion, to which the only two panelists to disagree were of the TMSS group. Reviewing the survey response data, we can conclude that panelists were sufficiently comfortable with the standard setting methods. The Angoff group indicated the highest levels of comfort with their procedure, on average and indicated by the frequencies in Table 3, and the Bookmark group indicated the lowest levels of

comfort, on average and likewise indicated in Table 3. The Bookmark and TMSS groups' lower overall comfort with their judgmental procedure relative to the Angoff group was echoed in group discussion. A common theme in both the Bookmark and TMSS groups' discussions was difficulty in placing cut-scores due to the diversity of the content (e.g., an item related to pulmonary medicine, followed by an item related to cardiac disorders, followed by an item related to medical ethics, etc.). Similarly, both groups expressed disbelief that the items were actually ordered by empirical difficulty.

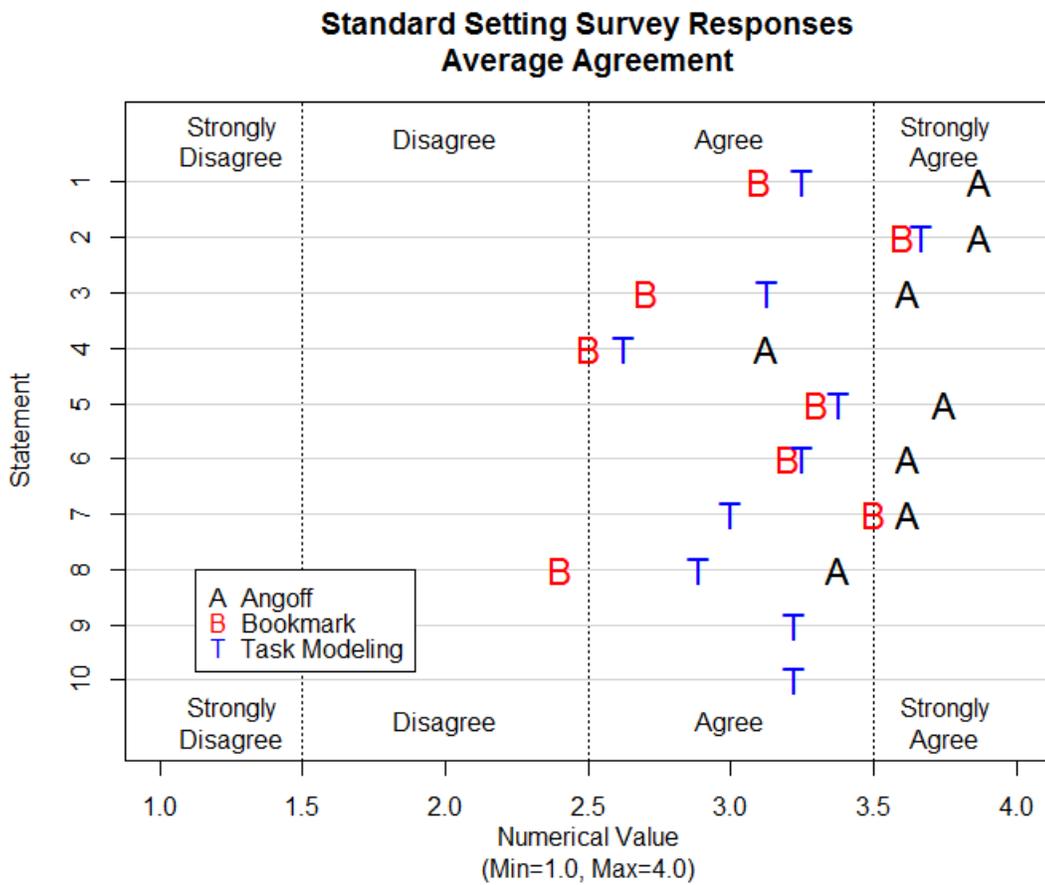


Figure 6. Average Agreement for Survey Responses

Final Cut-Score Rigor and Interpanelist Consistency

Outlined in the methods chapter, each group conducted two rounds of their prescribed standard setting method. As the first two research questions both pertained to the final cut-scores of each group, they will be discussed together here. Table 4, below, shows the recommended cut-score for each panelist in the first and second round, displayed in Rasch logits. Following Table 4, Figure 7 displays this information graphically. Looking at the information contained in Table 4, it is clear that the Angoff method resulted in the lowest, yet most consistent average cut-score between rounds (change of .01 between rounds, compared to changes of .08 and .17 for the Bookmark and TMSS, respectively). The Bookmark and TMSS cut-scores were found to be very similar, separated by only .01. In addition to the overall difference in cut-scores between the Angoff and Bookmark/TMSS groups as quantified by the final group cut-score, Table 4 and Figure 7 also demonstrate that six of the eight (6/8) panelists in the Angoff group had final cut-scores lower than the minimum cut-score recommended by any panelist in either the Bookmark or TMSS groups.

Unfortunately, the small sample size of each of the standard setting groups render parametric, inferential statistics ineffective (for example, the observed power of a one-way ANOVA on final cut-score by group is estimated at .39 with an eta-squared of .15); however, certain patterns do emerge in Table 4 and Figure 7. The Angoff group had noticeably more variability in their final cut-scores (and thereby less interpanelist consistency), with a standard deviation of .68 compared to very similar standard deviations of .14 for the Bookmark group and .12 for the TMSS group.

Table 4. Cut-Scores by Method and Round

Panelist	Angoff		Bookmark		Task Modeling	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
1	0.71	0.80	1.19	1.19	1.36	1.36
2	1.47	1.47	1.16	1.29	1.58	1.48
3	0.98	1.03	1.10	1.10	1.27	1.27
4	0.98	0.94	1.29	1.23	1.76	1.29
5	0.32	0.28	1.77	1.30	1.27	1.27
6	2.39	2.31	1.53	1.53	1.77	1.48
7	0.01	0.15	1.77	1.48	1.53	1.29
8	0.89	0.89	1.20	1.20	1.65	1.27
9	-	-	1.36	1.36	1.10	1.10
10	-	-	1.65	1.48	-	-
Mean	0.97	0.98	1.40	1.32	1.48	1.31
Median	0.94	0.92	1.33	1.30	1.53	1.29
Min	0.01	0.15	1.10	1.10	1.10	1.10
Max	2.39	2.31	1.77	1.53	1.77	1.48
SD	0.73	0.68	0.26	0.14	0.24	0.12
SE	0.26	0.24	0.08	0.05	0.08	0.04

The Angoff method resulted in the most consistent ratings within panelists between rounds. This is evident by the near perfect Spearman rank correlation of .99 for the Angoff group, and correlations of .78 for the Bookmark group and .67 for TMSS group. Figure 7 shows the overall pattern for each group, with the Angoff group staying more consistent, and the Bookmark and TMSS groups converging and decreasing in the second round. Despite more consistency between rounds, the Angoff group had only two of eight panelists make no changes to their item-level ratings in Round 2, after the group discussion. For the Bookmark group, five of the 10 (5/10) panelists remained with their original cut-scores. For the TMSS group, four of the nine (4/9) panelists remained with their original cut-scores.

Cut-Scores by Method and Round

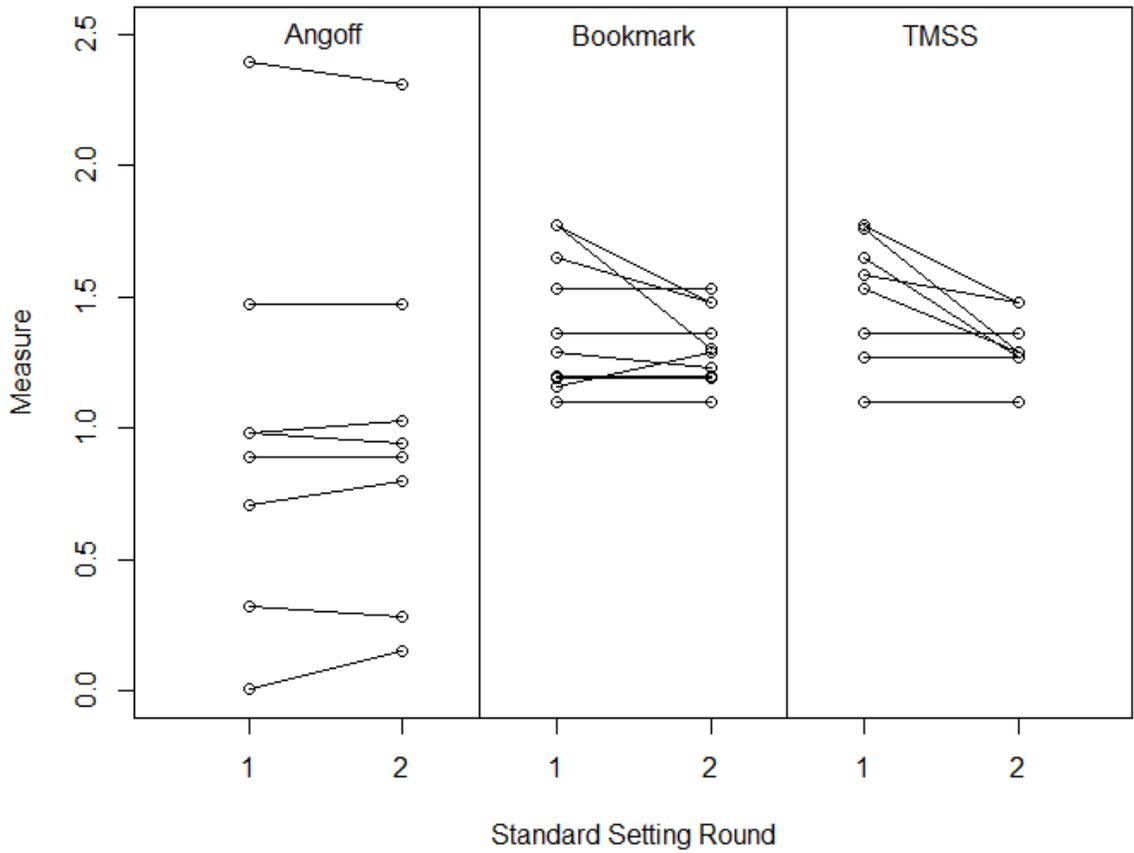


Figure 7. Graphical Display of Cut-Scores by Method and Round

Panelists' Confidence and Ability to Interpret the Scale

The third research question addressed panelists' confidence in their cut-scores, their perceived ability to differentiate between candidates of different abilities (i.e., candidates at different levels on the score scale), and if panelists were able to make statements regarding the KSAs of candidates based on their cut-scores. The hypothesis for this research question being that a greater focus on the KSAs of candidates in the TMSS would result in a better understanding by the panelist of their cut-score, and likewise a better grasp of the differences in KSAs manifest at different levels of the scale. This research question was evaluated using statements #4 and #8 on the final evaluation survey, as well as the final, open-ended item on the survey.

Looking back to Table 2, we can see that the Angoff group indicated higher levels of agreement with both of these statements than the Bookmark and TMSS groups. Only one (1/8) panelist in the Angoff group indicated disagreement to statement #4, and none (0/8) of the panelists disagreed with statement #8. In contrast, three panelists in each of the Bookmark and TMSS groups disagreed to some extent with statement #4. The Bookmark group appears to be the outlier for statement #8, as six of the ten (6/10) responses indicated disagreement to the statement, whereas seven of the nine (7/9) respondents for the TMSS group agreed to some extent with the statement. These findings suggest that the TMSS may have been an improvement over the Bookmark method, but the highest levels of confidence and understanding were observed for the Angoff group.

The final question on the evaluation survey asked panelists, if able, to provide statements regarding the competency of passing candidates, based on their recommended cut-score. As this was the last item on the survey prior to the request for comments on the back of the survey, some panelists used this space to provide general comments rather than respond to the directed topic. The general comments will be addressed momentarily, as the focus of the current discussion is the panelists' abilities to delineate the expected KSAs of passing candidates

Table 5, below, contains the counts of responses from each group when asked to provide statements as to the competencies of passing candidates based on their cut-score. Shown in the table, we have the total number of panelists in each group who responded to the survey question, the number of responses that were specific to KSAs (as opposed to general comments), and finally two example comments for each group- one that was classified as specific to KSAs and one that was classified as not specific to KSAs.

Table 5. Statements Regarding Competency of Passing Candidates

Group	Responses	Specific to KSAs	Specific Example	Non-specific Example
Angoff	4 of 8	3 of 4	Can safely care for [redacted] patients. Are knowledgeable in routine [redacted] illnesses/conditions. Know when a [redacted] patient is "sick" vs "non-sick".	They have a beyond basic knowledge of [subject matter].
Bookmark	9 of 10	3 of 9	Able to synthesize laboratory imaging data to formulate a differential diagnosis. Able to use recalled facts and apply to decision making.	Passing candidates are likely to achieve at least 60% of test questions correctly.
TMSS	5 of 9	4 of 5	Passing candidates are able to identify life threatening conditions that require immediate attention, identify abnormal conditions such as developmental delays (motor or speech) which require further evaluation, and are able to diagnose and manage common and less common (but not rare) disorders in [redacted] patients.	Given wide variety of questions/topics, I find it difficult to make statements regarding competency.

We can see the Bookmark group provided the most comments, with nine of the 10 (9/10) panelists providing information. However, only three of the nine (3/9) panelists providing information were able to provide statements specific to the KSAs of passing candidates. Similar counts of panelists provided comments in the Angoff and TMSS groups, and similar proportions of which were classified as specific to the KSAs of passing candidates. In considering the results from statements #4 and #8 in conjunction with the open-ended results, it is apparent that the Angoff group had the best

understanding of the scale (as defined by self-evaluated understanding and ability to articulate expectations), followed by the TMSS group.

Additional Results:

Hofstee Ratings

Described previously, a Hofstee exercise was conducted at the culmination of each of the three criterion-referenced standard settings used in this study. While the focus of this study was on criterion-referenced methods for setting standards, the Hofstee exercise may help to shed light on the comparability of each group's holistic view regarding the appropriate passing standard of the exam, as well as whether the passing standard they set in their judgmental task actually conforms to such a view. In Table 6 below, the final cut-scores by each group have been converted from Rasch logits to expected raw scores on the 100 items, and compared against the range of acceptable cut-scores as calculated in each group's Hofstee exercise.

Table 6. Raw Cut-Scores and Hofstee Acceptable Cut-Score Ranges

Group	Cut-score	Minimum Hofstee	Maximum Hofstee
Angoff	61	65	76
Bookmark	69	63	76
TMSS	69	65	77

Shown in Table 6, we can see that the three groups were consistent in their holistic judgments of minimum and maximum acceptable raw cut-scores on the 100 item form, providing evidence that, at the holistic level, the groups had similar conceptions as to the overall difficulty of the set of items. The minimum raw cut-scores ranged between

63 and 65 for the three groups, and the maximum raw cut-scores ranged between 76 and 77. For the Bookmark and TMSS groups, the recommended cut-score produced by their initial judgmental process fell in each group's range, and is strong evidence for the external validity of those cut-scores. However, the Angoff group's cut-score fell outside of their Hofstee exercise range, and does not provide external validity evidence in the form of conformity between standard setting methods.

CHAPTER V

SIGNIFICANCE, LIMITATIONS AND FUTURE RESEARCH

The current study sought to further incorporate criterion-referenced information, in the form of AE task models, into the criterion-referenced practice of standard setting. In doing so, the goal was to achieve a better understanding of the scale for panelists, remove noise from the process by focusing on the fundamental pieces of items that drive complexity, and assess how the incorporation of such information would affect the rigor and consistency of resulting cut-scores.

In conducting this study, two potential, immediate challenges were evident. First, task models had to be created for the items to be used in the standard setting. Specifically, since the items had already been created, task models needed to be “retrofit” to the items. Luecht, Burke, and Devore (2009) effectively fit task models to items in the certification/licensure arena, and the current study demonstrated that such an endeavor is feasible in the specific context of medical certification as well. There is also anecdotal evidence in this study to support SME buy-in and engagement when it comes to task modeling. The SMEs participating in the creation of the task models commented that they found value in the process, and found it more interesting than the item writing assignments to which they were accustomed.

Second, standard setting panelists participating in the TMSS had to be introduced to task models and prepared to use them as their unit of judgment in a training period prior to the standard setting procedure. Despite the fact that the training was only 30 minutes and covered multiple topics, panelists in the TMSS group indicated they understood task models conceptually and were comfortable with the standard setting process. This finding indicates that panelist training on the concepts/structures of AE may not be a barrier to incorporating such structures into more standard setting procedures in the future. In related research conducted by the author, a consistent warning from colleagues was to be wary of overloading panelists with too much information. The findings of this study imply that panelists are capable of comprehending principled assessment information while conducting a judgmental process. Following the conclusion of the study, one of the panelists who participated in the TMSS group remarked that they believed the panelists in the group were able to see the relationship between task models and setting passing standards.

Interestingly, and despite the level of comfort indicated by panelists with task models, five of the nine (5/9) panelists in the TMSS group indicated that they ultimately used the item, rather than task model, to make their final judgment. So while panelists can become comfortable with the concept of a task model rather quickly, evidence in this study suggested that some panelists will ultimately revert to using the item, if the item is available. Future research should investigate panelists' comfort with task models, and their overall comfort and confidence in the process, when items are not provided (or

provided in a more supporting role of the task model rather than side-by-side) to support the interpretation of the task model.

Given panelists understanding of task models, this opens the door for more prospective and confirmatory approaches to standard setting. The clear relationship between the criterion-referenced claims/expectations of passing candidates and the demands of task models set the stage for standards to be set early in the life cycle of an assessment, and guide the development of the item pool and exam form assembly. In such a way, standard setting in its current, post-hoc role would serve to validate or confirm the passing standard, as opposed to discovering it in an exploratory sense.

The testing organization used in this study anecdotally shared with the author their comfort historically, as well as their SMEs comfort historically, with the yes/no Angoff method. This study confirmed such a level of comfort, as demonstrated by the Angoff group's higher survey ratings across the board in comparison to the other two methods. While the panelists in the Angoff group had not previously participated in an Angoff standard setting, it is possible that they were familiar with the procedure prior to the standard setting, based on their involvement with the organization in other aspects of the exam, or through interactions with colleagues/committee members who had participated in standard settings.

Despite this level of comfort with the process, it was interesting to note that the Angoff group's ratings were the only ratings that did not fit with their holistic judgments made in the Hofstee exercise. Therefore, while the survey results provided strong procedural evidence for the Angoff method, the Bookmark and TMSS methods provided

stronger external validity evidence via consensus between multiple methods within each group. In regard to internal validity evidence, the Angoff ratings were also substantially more variable than that of the Bookmark and TMSS groups. One potential explanation for this finding could be that the Angoff panelists did not have a “status” indicator while making their ratings, in the sense that they did not have a running tally of “yes” ratings as they progressed through the items, unless they were to keep count internally. In contrast, the Bookmark and TMSS groups had a constant indicator of where they were setting their standard in the form of page numbers. However, this does not explain the sustained variation in Angoff ratings that was found to persist after the discussion between rounds, in which the distribution of cut-scores was reviewed.

Overall, the slightly superior procedural validity evidence compared to the Bookmark group, and the greater interpanelist consistency and external validity evidence compared to that of the Angoff group, warrant further attention for the inclusion of task models in standard setting. The Bookmark judgmental process was selected for adaptation to a task model-based approach in this study due to the philosophical focus on an ordered scale in the Bookmark method. The ordered progression fit nicely with the tenets of AE construct maps, evidence models, and task models. However, panelists remarked during the Bookmark and TMSS procedures that the diversity of the content made item-to-item and task-to-task distinctions difficult, and often questioned whether the booklet was truly ordered by empirical difficulties. For this reason, incorporating task models into an Angoff method, based on its historical comfort and comfort level confirmed in this study, may prove to be the most advantageous next step. Another

option would be attempt TMSS similar to its implementation in this study; however, do so within the different content domains that make up the larger blueprint of a testing program. This could prove challenging depending on the breadth and depth of the content blueprint (for example, the testing program in this study currently has over 30 unique content domains), but could help remedy the cognitive dissonance encountered by panelists in making their item-by-item judgments.

Despite previous findings regarding the comparability between standards set by the Angoff and Bookmark methods (see Peterson, Shulz, and Englehard, 2011), this study found substantial differences between the Angoff and the Bookmark-based judgmental processes. As this study used a yes/no Angoff, it may be assumed that panelists used a 50% probability of a correct response by the minimally-competent candidate as their breaking point between “yes” and “no” ratings. However, this would be an assumption, and future research should investigate whether or not panelists use such a criteria, or if they have a different mechanism by which they make yes/no ratings.

One of the obvious limitations of nearly all standard setting studies is a lack of sample size, subsequently leading to a lack of generalizability in findings. The operational constraints incumbent in conducting standard setting workshops render procuring large samples with which to set standards extremely difficult. With the ever-increasing capabilities of technology, virtual standard setting could help to alleviate the burden on resources of conducting large-scale standard setting. However, it is possible that the benefit of larger samples for standard setting through virtual communication comes at the expense of interpersonal interaction and discussion, which are valued

components of the standard setting process. Lastly, in regard to generalizing results, the contextual aspects of an individual standard setting (e.g., the dynamics of the test program, the construct of interest, varying stakeholder populations, etc.) often make generalizability to other areas difficult.

A final limitation in this study, and in standard setting studies in general, is the lack of a “true” standard against which to compare findings. As a result of the lack of an absolute standard against which to evaluate the results of a standard setting, the best evaluation of standard setting workshops lies in evaluating the validity of their outcomes using the frameworks discussed in this study. For this reason, the current study and studies like it, are crucial to critically evaluate standard setting methods and ways of improving the validity of outcomes reached via standard setting processes.

In conclusion, the purpose of this study was to investigate the effects of incorporating AE task models into standard setting, and compare the validity evidence for such a standard setting against two prominent, contemporary methods. The findings of this study suggest that the passing standard set by the TMSS group had the strongest collection of validity evidence of the three methods. This conclusion is substantiated by the greater procedural validity of the TMSS group compared to that of the Bookmark group, and the greater internal and external validity evidence compared to that of the Angoff group. Ultimately, the TMSS and Bookmark groups arrived at nearly identical passing standard recommendations; however, the greater procedural validity evidence is indicative of a greater understanding and comprehension of the process by the TMSS group, rendering the standard set by that method more defensible.

REFERENCES

- AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-597). Washington, D.C.: American Council on Education.
- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, predictive, and progressive approach to standard setting. In R. Lissitz (Ed.), *Assessing and Modeling Cognitive Development in School: Intellectual Growth and Standard Setting*. Maple Grove, MN: JAM Press.
- Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7.
- Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, 9, 215-235.
- Brennan, R. L. (1995). Standard setting from the perspective of generalizability theory. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments, Volume II* (pp. 269-290). Washington, D.C.: US Government Printing Office.
- Brennan, R. L. (2002). *Estimated standard error of a mean when there are only two observations* (CASMA Technical Note No. 1). Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessments.
- College Board (2013). Validated Achievement Level Descriptions for AP French, German, and Italian Language and Culture Exams. from <http://media.collegeboard.com/digitalServices/pdf/ap/apcentral/ap-world-languages-achievement-level-descriptions-dec-2013.pdf>
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1), 59-88.

- Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Westport, CT: Praeger.
- Cizek, G. J. (1996a). Setting passing scores. *Educational Measurement: Issues and Practice*, 15(2), 20–31. Cizek, G. J. (1996b). Standard setting guidelines. *Educational Measurement: Issues and Practice*, 15(1), 13–21, 22.
- Cizek, G. J. (2012). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed., pp. 3-14). New York: Routledge.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.
- 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 teacher examinations. *Journal of Educational Measurement*, 21(2), 113-129.
- Ebel, R. L. (1962). Content Standard Test Scores. *Educational and Psychological Measurement*, 22(1).
- Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. L. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed., pp. 79-106). New York: Routledge.
- Ferrara, S., & Lewis, D. M. (2012). The Item-Descriptor (ID) Matching Method. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed., pp. 255-282). New York: Routledge.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological review*, 19(2), 139-150.
- Hambleton, R. K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 15, 277–290.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.
- Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On Education Testing* (pp. 109-127). San Francisco: Jossey-Bass.

- Huynh, H. (1998). On score locations of binary and partial credit items and their application to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, 23, 35-56.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 353-366.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). Washington, D. C.: American Council on Education.
- Jaeger, R. M., J. Cole, D. M. Irwin, and D. J. Pratto. (1980). An Interactive Structure Judgment Process for Setting Passing Scores on Competency Tests Applied to the North Carolina High School Competency Tests in Reading and Mathematics. Greensboro: Center for Education Research and Evaluation, University of North Carolina at Greensboro.
- Kaftandjieva, F. (2010) *Methods for setting ut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem, The Netherlands: CITO.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461.
- Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Standard Setting: Concepts, Methods, and Perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.
- Kane, M. T. (1999). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences in Education*, 4, 195-207.
- Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: Praeger.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.). New York: Springer-Verlag.
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*.

- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). *The bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the meeting of the 1998 National Council on Measurement in Education, San Diego, CA.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures using behaviorial anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.
- Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Shulz, E. M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed., pp. 225-253). New York: Routledge.
- Linn, R. L., Madaus, G. F., & Pedulla, J. T. (1982). Minimum competency testing: Cautions on the state of the art. *American Journal of Education*, *91*(1), 1-35.
- Loomis, S., C. (2012). Selecting and training standard setting participants: State of the art policies and procedures. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed., pp. 107-136). New York: Routledge.
- Luecht, R. M. (2006). *Engineering the test: From principled item design to automated test assembly*. Invited paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.
- Luecht, R. M. (2013a). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*.
- Luecht, R. M. (2013b). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 59-76). New York, NY: Routledge.
- Luecht, R. M., Burke, M., & Devore, R. (2009). Task modeling of complex computer-based performance exercises. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.
- Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, *5*, 265-283.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure Testing: Purposes, Procedures, and Practices* (pp. 219-252). Lincoln, NE: Burors Institute of Mental Measurements.
- Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I* (pp. 73-85). Washington, D. C.: Council of Chief State School Officers.
- Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (ed.), *Educational measurement* (4th ed., pp. 257-305). Westport, CT: American Council on Education and Praeger.
- Mislevy, R. J.; & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.
- National Academy of Sciences. (2005). *Measuring Literacy: Performance Levels for Adults, interim report*. Washington, D. C.
- No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).
- Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement*, 24(1), 56-64.
- Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, 10(1), 39-59.
- Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Koller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation*, 35, 95-101.

- Peterson, C. H., Schulz, E. M., & Engelhard, G. (2011). Reliability and validity of Bookmark-based methods for standard setting: Comparisons to Angoff-based methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 30(2), 3-14.
- Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-262). New York: Macmillan.
- Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and Yes/No standard setting methods. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed., pp. 181-199). New York: Routledge.
- Plake, B. S., Huff, K., & Reshetar, R. (2010). Evidence-Centered Assessment Design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education*, 23(4), 342-357.
- Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Standard Setting: Concepts, Methods, and Perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum.
- Reckase, M. D., & Bay, L. (1999). *Comparing two methods for collecting test-based judgments*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, 4, 447-467.
- Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels. In *Joint conference on standard setting for large-scale assessments. Vol. 2. Proceedings* (pp. 143-160). Washington, DC: U.S. Government Printing Office.
- Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting Performance Standards for Student Achievement*. Stanford, CA: National Academy of Education.
- Thorndike, R. L. (Ed.). (1971). *Educational Measurement* (2nd ed.). Washington, D. C.: National Council on Education.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(17), 433-451.

- Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement, 40*(3), 231-253.
- Williams, N. J., & Schulz, E. M. (2005). *An investigation of response probability (RP) values used in standard setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.
- Wilson, M. (2005). *Constructing Measures*. Mahway, NJ: Erlbaum.
- Wyse, A. E. (2013). Construct maps as a foundation for standard setting. *Measurement: Interdisciplinary Research and Perspectives, 11*(4), 139-170.
- Yen, W M., & Ferrara, S. (1997) The Maryland School Performance Assessment Program: Performance assessment with psychometric quality suitable for high stakes usage. *Educational and Psychological Measurement, 57*, 60–84.
- Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Standard Setting: Concepts, Methods, and Perspectives* (pp. 19-51). Mahwah, NJ: Erlbaum.
- Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the basic skills assessment tests*. Princeton, NJ: Educational Testing Service.

APPENDIX A

BOOKMARK AND TMSS RATINGS SHEETS

Bookmark Standard Setting Procedure

Please enter your panelist number: _____

Directions. Below, please enter the page in your item book at which point you judge the probability of a correct response by the borderline candidate to drop below $2/3$.

(i.e., at which point the borderline candidate no longer has a $2/3$ chance of responding correctly to the subsequent items)

Round 1: _____

Round 2: _____

Task Model Standard Setting

Please enter your panelist number: _____

Directions. Below, please enter the page in your item book at which point you judge the probability of a correct response by the borderline candidate to drop below $2/3$.

(i.e., at which point the borderline candidate no longer has a $2/3$ chance of responding correctly to the subsequent items)

Round 1: _____

Round 2: _____

APPENDIX B

SCREENSHOT OF QUALTRICS SURVEY FOR ANGOFF RATINGS

Directions. Each of the following item numbers corresponds to an item in your item book.

For each item, please respond with a "yes" or "no" as to whether the borderline candidate will answer this question correctly.

Item1.

Yes

No

Item2.

Yes

No

APPENDIX C

ANGOFF AND BOOKMARK SURVEY

Please enter your panelist number: _____

Please read each of the following statements carefully. Place a check or “X” in the box that best reflects your level of agreement with each statement.

*SD = Strongly Disagree, D = Disagree, A = Agree, SA = Strongly Agree

#	Statement	SD	D	A	SA
1	The training for this standard setting method was clear to me.				
2	I understood the goal of this standard setting.				
3	I am confident I applied the standard setting method appropriately.				
4	I am confident my final cut-score appropriately distinguishes between competent candidates and candidates who are not competent to practice.				
5	The discussion between rounds was helpful to me in completing my second round of ratings.				
6	I found the distribution of recommended cut-scores (minimum, maximum, and mean/median) following round one useful.				
7	The information showing the distribution of examination scores was helpful.				
8	I am confident in my ability to distinguish between the knowledge, skills, and abilities of candidates who achieve different scores (e.g., the difference between a candidate who scores 40 versus a candidate who scores 60).				

Based on your recommended cut-score and the items you reviewed, what statements can be made regarding the competency of passing candidates? (If able, please provide three example statements)

PLEASE PROVIDE ANY ADDITIONAL COMMENTS ON THE REVERSE SIDE

APPENDIX D

TMSS SURVEY

Please enter your panelist number: _____

Please read each of the following statements carefully. Place a check or “X” in the box that best reflects your level of agreement with each statement.

*SD = Strongly Disagree, D = Disagree, A = Agree, SA = Strongly Agree

#	Statement	SD	D	A	SA
1	The training for this standard setting method was clear to me.				
2	I understood the goal of this standard setting.				
3	I am confident I applied the standard setting method appropriately.				
4	I am confident my final cut-score appropriately distinguishes between competent candidates and candidates who are not competent to practice.				
5	The discussion between rounds was helpful to me in completing my second round of ratings.				
6	I found the distribution of recommended cut-scores (minimum, maximum, and mean/median) following round one useful.				
7	The information showing the distribution of examination scores was helpful.				
8	I am confident in my ability to distinguish between the knowledge, skills, and abilities of candidates who achieve different scores (e.g., the difference between a candidate who scores 40 versus a candidate who scores 60).				
9	I understand the concept of a task model as it was defined in this standard setting.				

10	I understood the individual task models that were presented in this standard setting.				
----	---	--	--	--	--

Each page in the item book contained a task model and an item. Please indicate whether the task model or the item played a larger role in your decision for the majority of the item book pages.

Task model
 Item
 Task model and item played an equal role

Based on your recommended cut-score and the items you reviewed, what statements can be made regarding the competency of passing candidates? (If able, please provide three example statements)

PLEASE PROVIDE ANY ADDITIONAL COMMENTS ON THE REVERSE SIDE