A Comparison of the Angoff and Item Mapping Standard Setting Methods for a Certification Examination

BY

SARAH D. SCHNABEL
B.S., University of Illinois at Champaign-Urbana, 2006
M.Ed., University of Illinois at Chicago, 2012

THESIS

Submitted as partial fulfillment of the requirements for the degree of Doctor of Philosophy in Educational Psychology in the Graduate College of the University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:

Everett Smith, Ph.D, Chair and Advisor
Maria Incrocci, R.Ph., Ph.D., National Association of Boards of Pharmacy
Kimberly Lawless, Ph.D.
Nicole Risk, Ph.D., American Medical Technologists
Yue Yin, Ph.D.
Acknowledgements

Thank you to my advisor, Ev, for his support and guidance during my time as a graduate student. Despite always working on a million things at once, you never missed responding to a single e-mail or question I had. To my committee members: Maria, Kim, Nicole, and Yue. It has been a pleasure learning from all of you, and I am extremely fortunate to have so much expertise on a single committee.

In addition to my committee members, I’d like to thank my colleagues at the American Board of Ophthalmology. First, to the volunteers who made my study possible: thank you. Perhaps I’m kidding myself, but it seems like you all love standard setting as much as I do sometimes! A special thanks to my leader, Beth Ann, who supported my research despite all the bumps in the road. You often remind me that I have an important voice in the field of assessment, and I will forever be grateful.

Lastly, to my family and friends - thank you for sticking by my side all these years despite my always being stretched incredibly thin. I am always inspired by my mother, Carletta, who is always able to convince me that I am capable of anything. And to my wonderful husband Brandon: thank you for everything. Your work ethic inspires me every day. You always remind me that things just aren’t that bad. For all the slack you’ve been picking up for the past seven and a half years – thank you. I’ll start helping with the dishes now. And you’ll never have to hear me say “I’m almost done with my dissertation, I just need one more hour” and disappear into the basement ever again!
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>A. Purpose of the Study</td>
<td>3</td>
</tr>
<tr>
<td>B. Research Question</td>
<td>4</td>
</tr>
<tr>
<td>C. Approach of the Study</td>
<td>4</td>
</tr>
<tr>
<td>D. Significance of the Study</td>
<td>5</td>
</tr>
<tr>
<td>II. REVIEW OF THE LITERATURE</td>
<td>7</td>
</tr>
<tr>
<td>A. Establishing a Passing Standard</td>
<td>7</td>
</tr>
<tr>
<td>B. Standard Validation</td>
<td>9</td>
</tr>
<tr>
<td>C. Existing Standard Setting Method Features and Classification</td>
<td>11</td>
</tr>
<tr>
<td>1. Method Classification</td>
<td>12</td>
</tr>
<tr>
<td>D. A History of Standard Setting Methodology</td>
<td>13</td>
</tr>
<tr>
<td>1. Nedelsky</td>
<td>13</td>
</tr>
<tr>
<td>2. Angoff</td>
<td>13</td>
</tr>
<tr>
<td>3. Ebel</td>
<td>14</td>
</tr>
<tr>
<td>4. Jaeger</td>
<td>15</td>
</tr>
<tr>
<td>5. Compromise methods</td>
<td>15</td>
</tr>
<tr>
<td>6. Bookmark method</td>
<td>16</td>
</tr>
<tr>
<td>7. Item mapping</td>
<td>17</td>
</tr>
<tr>
<td>E. Current Standard Setting Practices</td>
<td>18</td>
</tr>
<tr>
<td>1. Empirical performance data, feedback, and iterations</td>
<td>19</td>
</tr>
<tr>
<td>2. Adjustments</td>
<td>21</td>
</tr>
<tr>
<td>F. Criticisms of the MCE Methods</td>
<td>21</td>
</tr>
<tr>
<td>1. Inter-rater reliability</td>
<td>22</td>
</tr>
<tr>
<td>2. Panelist confidence</td>
<td>23</td>
</tr>
<tr>
<td>3. Empirical performance data, feedback, and iterations</td>
<td>23</td>
</tr>
<tr>
<td>4. Minimal competence</td>
<td>23</td>
</tr>
<tr>
<td>G. Method Comparisons</td>
<td>25</td>
</tr>
<tr>
<td>H. Summary of Literature</td>
<td>26</td>
</tr>
<tr>
<td>III. METHOD</td>
<td>28</td>
</tr>
<tr>
<td>A. Research Questions</td>
<td>28</td>
</tr>
<tr>
<td>B. Examination Properties</td>
<td>28</td>
</tr>
<tr>
<td>C. Study Design</td>
<td>31</td>
</tr>
<tr>
<td>1. Items</td>
<td>31</td>
</tr>
<tr>
<td>2. Participants (panelists)</td>
<td>35</td>
</tr>
<tr>
<td>D. Standard Setting Procedures</td>
<td>36</td>
</tr>
<tr>
<td>1. Schedule</td>
<td>36</td>
</tr>
<tr>
<td>2. Initial training</td>
<td>37</td>
</tr>
<tr>
<td>3. Method 1 (Angoff)</td>
<td>39</td>
</tr>
</tbody>
</table>
### TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.</td>
<td>Method 2 (item mapping)</td>
</tr>
<tr>
<td>5.</td>
<td>Evaluation</td>
</tr>
<tr>
<td>6.</td>
<td>Surveys</td>
</tr>
<tr>
<td>E.</td>
<td>Analysis Approach</td>
</tr>
<tr>
<td>1.</td>
<td>Initial data analysis</td>
</tr>
<tr>
<td>2.</td>
<td>Procedural validity evaluation</td>
</tr>
<tr>
<td>3.</td>
<td>Internal validity evaluation</td>
</tr>
<tr>
<td>4.</td>
<td>External validity evaluation</td>
</tr>
<tr>
<td>a.</td>
<td>Assessing reasonableness</td>
</tr>
<tr>
<td>b.</td>
<td>Hofstee method comparison</td>
</tr>
<tr>
<td>F.</td>
<td>Summary of Methodology</td>
</tr>
<tr>
<td>IV.</td>
<td>RESULTS</td>
</tr>
<tr>
<td>A.</td>
<td>Participants</td>
</tr>
<tr>
<td>B.</td>
<td>Ratings and Cut Scores</td>
</tr>
<tr>
<td>1.</td>
<td>Angoff</td>
</tr>
<tr>
<td>2.</td>
<td>Item mapping</td>
</tr>
<tr>
<td>3.</td>
<td>Cut scores</td>
</tr>
<tr>
<td>C.</td>
<td>Intra-Panelist Consistency</td>
</tr>
<tr>
<td>D.</td>
<td>Inter-Panelist Consistency</td>
</tr>
<tr>
<td>E.</td>
<td>Surveys</td>
</tr>
<tr>
<td>1.</td>
<td>Initial training survey</td>
</tr>
<tr>
<td>2.</td>
<td>Angoff survey</td>
</tr>
<tr>
<td>3.</td>
<td>Item mapping survey</td>
</tr>
<tr>
<td>4.</td>
<td>Final results survey</td>
</tr>
<tr>
<td>5.</td>
<td>Post-rating evaluation conference call</td>
</tr>
<tr>
<td>V.</td>
<td>DISCUSSION</td>
</tr>
<tr>
<td>A.</td>
<td>Procedural Validity</td>
</tr>
<tr>
<td>1.</td>
<td>Panelist representation</td>
</tr>
<tr>
<td>2.</td>
<td>Procedural understanding</td>
</tr>
<tr>
<td>B.</td>
<td>Internal Validity</td>
</tr>
<tr>
<td>1.</td>
<td>Intra-panelist consistency</td>
</tr>
<tr>
<td>2.</td>
<td>Inter-panelist consistency</td>
</tr>
<tr>
<td>C.</td>
<td>External Validity</td>
</tr>
<tr>
<td>1.</td>
<td>Hofstee method comparison</td>
</tr>
<tr>
<td>2.</td>
<td>Standard reasonableness</td>
</tr>
<tr>
<td>D.</td>
<td>Summary</td>
</tr>
<tr>
<td>E.</td>
<td>Significance and Practical Implications</td>
</tr>
</tbody>
</table>

REFERENCES ............................................................................................................. 82
TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPENDICES</td>
<td>97</td>
</tr>
<tr>
<td>Appendix A</td>
<td>93</td>
</tr>
<tr>
<td>Appendix B</td>
<td>94</td>
</tr>
<tr>
<td>Appendix C</td>
<td>95</td>
</tr>
<tr>
<td>Appendix D</td>
<td>96</td>
</tr>
<tr>
<td>Appendix E</td>
<td>97</td>
</tr>
<tr>
<td>Appendix F</td>
<td>98</td>
</tr>
<tr>
<td>Appendix G</td>
<td>100</td>
</tr>
<tr>
<td>VITA</td>
<td>114</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>EXAMINATION CONTENT AND PSYCHOMETRIC PROPERTIES</td>
</tr>
<tr>
<td>II.</td>
<td>EXAMINATION FORM ITEM PERFORMANCE</td>
</tr>
<tr>
<td>III.</td>
<td>EXAMINATION FORM PERSON PERFORMANCE</td>
</tr>
<tr>
<td>IV.</td>
<td>PANELIST COMPOSITION BY GROUP</td>
</tr>
<tr>
<td>V.</td>
<td>ANGOFF RATINGS BY PANELIST</td>
</tr>
<tr>
<td>VI.</td>
<td>ITEM MAPPING RATINGS AND CUT SCORE SELECTION BY PANELIST</td>
</tr>
<tr>
<td>VII.</td>
<td>CUT SCORES BY GROUP AND METHOD</td>
</tr>
<tr>
<td>VIII.</td>
<td>HISTORICAL CUT SCORES FOR EXAMINATION</td>
</tr>
<tr>
<td>IX.</td>
<td>INTRA-PANELIST CONSISTENCY</td>
</tr>
<tr>
<td>X.</td>
<td>INTER-PANELIST CONSISTENCY</td>
</tr>
<tr>
<td>XI.</td>
<td>POST-INITIAL TRAINING SURVEY</td>
</tr>
<tr>
<td>XII.</td>
<td>POST ANGOFF SURVEY</td>
</tr>
<tr>
<td>XIII.</td>
<td>POST ITEM MAPPING SURVEY</td>
</tr>
<tr>
<td>XIV.</td>
<td>HOFSTEE SURVEY RESULTS</td>
</tr>
<tr>
<td>XV.</td>
<td>FINAL CUT SCORE SURVEY RESULTS</td>
</tr>
<tr>
<td>XVI.</td>
<td>KEY FINDINGS FROM POST-EVALUATION CONFERENCE CALL DISCUSSION</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Frequency distribution of raw scores on examination</td>
<td>30</td>
</tr>
<tr>
<td>2. Standard setting activities schedule</td>
<td>37</td>
</tr>
<tr>
<td>3. Hofstee method plot</td>
<td>65</td>
</tr>
</tbody>
</table>
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>IQR</td>
<td>Interquartile Range</td>
</tr>
<tr>
<td>MCE</td>
<td>Minimally-Competent Examinee</td>
</tr>
<tr>
<td>MCI</td>
<td>Multiple-Choice Item</td>
</tr>
<tr>
<td>NAEP</td>
<td>National Assessment for Educational Progress</td>
</tr>
<tr>
<td>RIDP</td>
<td>Rasch Item Difficulty Parameter</td>
</tr>
<tr>
<td>RP</td>
<td>Response Probability</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SEJ</td>
<td>Standard Error of Judgment</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard Error of Measurement</td>
</tr>
</tbody>
</table>
SUMMARY

Standard setting is the process of categorizing performance levels on an examination, often into two groups: pass and fail. It is essential that for criterion-referenced examinations, standards are set in a deliberate and meaningful way. In licensure and certification, individuals who pass an examination are granted the ability to practice or recognized as having met a high standard of knowledge. It is essential to develop an examination that measures the knowledge, skills, and abilities needed to practice competently. One component of that process is selection of a passing standard or cut score that reflects the minimum amount of knowledge necessary for practice at that level.

Standards are typically set by a group of stakeholders or subject-matter experts (often individuals that have been granted the license or certificate already). These individuals (panelists or participants) convene to establish a standard for an examination. Many standard setting methods exist, among them methods that require panelists to review the individual test questions (items) and make judgments about them. Those judgments (ratings) are then compiled and transformed into a cut score.

This research evaluates two standard setting methods: the Angoff method and the item mapping method. Although similar, these methods differ in the way panelists make judgments about items. The Angoff method variation used in this study requires panelists to estimate the likelihood that a minimally-competent examinee would answer an item correctly. The item mapping method makes use of a graphical representation of items and encourages group discussion. Using the item mapping method, panelists answer “yes” or “no” to the question “would a minimally-competent examinee answer this item correctly?”
SUMMARY (continued)

In this study, the two methods were compared with respect to procedural validity, internal validity, and external validity. The evidence for procedural validity for both methods was strong. Neither method was significantly more internally consistent, but the Angoff method resulted in higher intra-panelist consistency (stronger correlations between ratings and empirical difficulty). External validity is established in part through evaluation of the reasonableness of the cut score. The panelists did not agree on which method produced a more reasonable cut score, but they preferred the process of the Angoff method. Both methods had strengths and weaknesses which are explored in detail in the research.

This study demonstrates the importance of evaluating standard setting methods in a variety of ways. Additional research on these and other methods should include both quantitative and qualitative analyses. Because standard setting is a critical component of the examination process, it is imperative that standards are set in a meaningful way.
I. INTRODUCTION

The purpose of professional certification is to credential a group of individuals who have achieved a certain level of knowledge, skill, or ability in their field of practice. Generally, some combination of pre-requisite training or experience and a passing score on an examination is necessary to obtain certification. Ensuring that requirements for certification (and thus the pool of credentialed individuals) meet this standard is essential to maintaining the certification’s value. While all certification requirements contribute to its value, achieving a passing score on an examination remains a contentious and scrutinized requirement. Quantifying “what” and “how much” an individual needs to know (passing standard) and translating it into an actual, measurable score (cut score) is called standard setting. Cizek (1993) provides a more operational definition: standard setting is “the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance” (p. 100). Sound standard setting is essential in certification examinations; that is, the passing standard must reflect the minimum amount of knowledge needed to uphold the value of that certification. This idea, the validity of the interpretation of the passing standard, must be evaluated for all high-stakes examinations. Unfortunately, no one standard setting method is most appropriate for all certification examinations.

Evaluating the merits of different standard setting methodologies is a necessary first step for any examination program intending to implement (a) cut score(s). This research will focus on standard setting methods appropriate for a high-stakes certification examination consisting of multiple-choice items (MCIs).
Many standard setting methods (and variations of those methods) have been proposed and evaluated. One way to classify methods is by separating them into examinee-centered methods and examination-centered methods. The examinee-centered methods primarily rely upon an individual’s performance on an examination as the basis for discussion and judgment, whereas examination-centered methods focus on individual examinations or examination items (Jaeger, 1989). Another classification scheme separates methods into norm-referenced and criterion-referenced methods; the former is based upon a number or percentage of individuals pre-determined to pass while the latter is based upon meeting a criterion that is not affected by how other examinees perform (i.e., everyone can pass the examination if they have the necessary knowledge). Examination-centered, criterion-referenced methods are used most frequently in high-stakes examinations.

The standard setting process typically occurs in two steps. First, a performance standard is conceptualized. Then, that standard is translated into a cut score or passing score. The examination-centered, criterion-referenced methods typically approach the first step in a similar way; a group of subject matter experts (“judges” or “raters” or “panelists”) is gathered to define the minimal acceptable level of knowledge or competence for individuals in the population of interest. A standard setting facilitator enables discussion among panelists, which produces a list of the knowledge, skills, and abilities that need to be present to achieve a passing score (often referred to as a definition of minimal competence).

The approaches to translating that passing standard to a cut score are varied. The Angoff (1971) and Ebel (1972) methods rely on panelists to estimate the number of minimally-competent examinees (MCEs) who would answer an item correctly. The Nedelsky (1954) method requires panelists to identify the distractors in each item that an MCE would be able to
identify as incorrect. Once panelists’ response data are tallied, a second round of judging items may occur after the provision of some information. Often this information includes impact data or measures of panelist (inter-rater) agreement. Sometimes, a compromise standard setting method is used (Beuk, 1984; Hofstee, 1983), which utilizes additional normative information to set a standard.

Wang explored some of the shortcomings of the Angoff, or modified-Angoff, standard setting approach and presented an alternative procedure. Wang noted that standard-setting judges (panelists) are unable to accurately predict the performance of borderline examinees on items, which is the central task required in the Angoff (and many other MCE) methods. The item mapping procedure she presented, which utilizes a Rasch measurement model, simplifies the prediction. With item mapping, panelists see items in a histogram. The items appear in order from easiest to hardest based upon their Rasch item difficulty parameter (RIDP) estimates. Then, the group reviews items one at a time, starting with the easiest item with each subsequent item being progressively more difficult. After reviewing an item, the panelists determine whether 50% or more of the borderline candidates would answer that item correctly. The process is repeated until the panelists reach agreement on what item (or items) would result in 50% of borderline examinees answering correctly. The item(s) location in the histogram corresponds with a certain ability level. This ability level can then be translated into a cut score on the examination (Wang, 2003; 2009).

A. **Purpose of the Study**

In this study, I examined the Angoff and item mapping standard setting methods to determine which method or aspect(s) of each better support interpretation validity of pass/fail scores in a high-stakes MCI medical recertification examination. Using one form of an
examination, I facilitated standard setting sessions during which panelists set standards using each method. I then compared the recommended cut scores, ratings, and experiences of the panelists.

B. **Research Question**

This research intends to answer the question: Which standard setting method (item mapping or Angoff) leads to a more appropriate cut score on this high-stakes MCI exam?

Each standard setting method was evaluated with respect to procedural validity, internal validity, and external validity. To evaluate procedural validity, both methods were evaluated regarding their use of a diverse and representative panel and the indication of panelists that they understood the rating tasks they were asked to engage in. Secondly, to evaluate internal validity, intra-panelist consistency and inter-panelist consistency were calculated and compared for each method. Lastly, external validity was evaluated through surveying the panelists and convening a feedback discussion to determine which method’s cut score was perceived as more reasonable.

C. **Approach of the Study**

I selected a medical recertification examination for this comparative study. The examination consists of 50 operational MCIs. One half of the panelists rated these items using first the Angoff method and then the item mapping method. The other half rated the items using the item mapping method and then the Angoff method. The panelists were split into groups to balance gender, geographical location, subspecialty area, type of practice, and amount of time since initial certification. The standard setting exercises took place remotely. After the standard setting rating exercises, a conference call was convened to gather feedback from the panelists on both methods and the resulting cut scores. Panelists were asked which method’s judgments they were more confident making and which resulting cut score was more reasonable in their opinion.
The Angoff method requires panelists to conceptualize an MCE or a group of MCEs. Then, panelists estimate the likelihood that an MCE (or a group of MCEs) would answer each exam item correctly (from 0% to 100%).

After making an initial item rating, panelists are provided with item performance data, and then they rate the item a second time (round 2). The average round 2 panelist rating for each item is averaged to obtain a recommended cut score (percentage correct on the entire examination). This method was selected both because it is a common application of the Angoff method and because it is the method that has been used to set standards on this examination in the past.

The item mapping method requires panelists to review examination items that have been organized graphically to show relative difficulty. The judgment made on items is similar to that of the Angoff method, but it requires a dichotomous (“yes” or “no”) response to the question: “does an MCE have at least a 50% chance of answering this item correctly?” The recommended cut score corresponds to the average item difficulty for the column of items in the histogram for which the panelists agreed that the MCE had a 50% chance of answering the items correctly.

D. **Significance of the Study**

This study does not establish a “gold standard” for standard setting; however, given the importance of high-stakes criterion-referenced examinations, the need for a method to support interpretation validity of examination scores is vital. From the research already presented, it is clear that more research would greatly benefit examination developers and, indirectly, the protection of the public. Results from this study will assist the research community in exploring

---

1 Angoff (1971) proposed two variations of his method: 1) estimating whether an MCE would answer each exam item correctly; and 2) estimating the likelihood that an MCE (or a group of MCEs) would answer each exam item correctly (from 0% to 100%). The latter, sometimes referred to as the “modified-Angoff method,” is often considered the more widely-used variation (Zieky, 2012). The 0-100 variation is used in this study but is typically referred to simply as the “Angoff method.”
the potential of the lesser-known item mapping method. The strengths and weaknesses of the Angoff and item mapping methods are evaluated, which will allow researchers to select methods or method variations that best suit their needs. This study has the potential to influence examination research in the larger educational research community. Standard setting is approached similarly in many high-stakes criterion-referenced examinations, so the findings could influence standard setting in non-medical fields (e.g., teacher certification) and education.
II. REVIEW OF THE LITERATURE

Standard setting is the process of establishing a cut score or passing standard on an examination. In some cases, more than one cut score or passing standard may be needed to differentiate among different levels of performance or mastery. In contexts where an examination is used to determine whether an individual should be licensed, promoted, or graduate, only one cut score is needed. The cut score separates individuals into two categories: pass and fail. Individuals who pass an examination are credentialed with a license or certificate that designates them as legally allowed to practice, or designates them as providing more “advanced or specialized skills” (Schmidt, 1995, p. 18). In medical licensure/certification, identifying appropriate individuals to whom to grant a certificate or license impacts public health. Licensure is typically defined as mandatory, while certification is voluntary (Schmidt, 1995); however, “…licensure bestows the status of legal independent practitioner, for practical reasons certification may become a type of de facto license” (Clauser, Margolis, & Case, 2006, p. 701). For these reasons, great care needs to be taken when establishing a cut score for licensure or certification examinations due to the potential consequences resulting from score use.

A. Establishing a Passing Standard

Establishing a passing standard and corresponding cut score on an examination is necessary for pass/fail criterion-referenced examinations. A criterion-referenced measure is one that assesses an individual’s performance with respect to a continuum of knowledge, as opposed to in relation to the performance of others (Glaser, 1963). Hambleton, Swaminathan, Algina, and Coulson (1978) noted two uses for criterion-referenced examinations; “estimation of examinee domain scores and allocation of examinees to mastery states” (p. 5). It has since been argued that
depending upon the purpose of the examination, standards are not always needed for criterion-referenced examinations (Shepard, 1980); however, when a pass/fail decision must be made, a standard is essential. Throughout this paper, the term “passing standard” will refer to the “conceptual version of the desired level of competence” (Kane, 1994, p. 426), and “cut score” will refer to the corresponding examination-specific passing standard (usually described in terms of number of items or percentage of items needed to answer correctly in order to pass).

Nedelsky (1954) advocated for developing absolute standards, which are based upon quantification of some amount of mastery needed. The alternative, a relative standard, is developed by establishing a set percentage or number of individuals who will pass (e.g., everyone scoring higher than one standard deviation below the mean). This alternative practice results in failing some individuals regardless of their knowledge. It logically follows that, if a test is meant to establish whether an individual has mastered content, the performance of other examinees should not affect an individual’s score, and, thus, implementing a relative standard is not sound practice. The litany of standard setting research conducted since Nedelsky’s 1954 research largely advocates setting absolute standards in contexts where there is not a predetermined number of candidates who will advance, which is often the case in licensure and certification (AERA, APA & NCME, 2014; Norcini & Shea, 1997).

The question that arises is: “How do we establish an absolute standard on a criterion-referenced examination?” Glass (1978) contended that development of an absolute standard that is representative of mastery is not possible. He stated, “The language of performance standards is pseudoquantification, a meaningless application of numbers to a question not prepared for quantitative analysis.” (p. 238). Standard setting usually occurs in two steps, although not always discretely; first, defining minimal competence (conceptualize a passing standard) and second,
translating that definition into a cut score. Glass (1978) argued that, because a minimal acceptable level of performance cannot be specified, subsequent steps (quantification) are meaningless. Glass’s contribution is clear: validating the existence of a minimal acceptable level of performance is needed before we evaluate methods of quantification. This concept, whether “minimal competence” can be defined and quantified, will be revisited later in the paper.

While Glass concluded that standard setting is arbitrary, other researchers have argued that while it is subjective and judgmental, it is not necessarily arbitrary (Cizek, 1993; Shepard, 1980). Popham (1978) suggested that it is possible to arrive at reasonable standards despite inherent subjectivity. While the subjective nature of the task undeniably adds to the complexity, it is clear that decisions will be made based upon examination results whether or not a standard has been established. Thus, “[t]here is a need for standard setting… simply because decisions must be made” (Cizek, 2006, p. 227).

If we accept that standards must be established despite the subjectivity, we must determine methods for developing them that will support the intended argument: that individuals who pass the examination have mastered the necessary knowledge domain, and individuals who fail the examination have not.

B. Standard Validation

Validity is the “degree to which evidence and theory support the interpretations of test scores for proposed uses of test” (AERA, APA & NCME, 2014, p. 11). Validity is a property of the intended use of test scores and making an argument for or against validity requires an evaluation of all evidence in support of and against the intended use of scores (Cronbach, 1988; Messick, 1989). “The basic issue of validity is whether the statements and decisions derived from test scores are appropriate” (Kane, 1994, p. 134).
What is clearly difficult about evaluating the validity of a pass/fail examination score interpretation is that there is little (if anything) that can be done to confirm “correctness.” There is no examination validation activity that will prove that the dichotomization of examinees into two groups by means of a standard is accurate (Jaeger, 1990). This is true for two reasons. First, dichotomization is judgmentally based, and individuals who score just above and below the cut score will likely have similar ability levels (Jaeger, 1990). Second, it is not possible for a test administrator to know that a cut score properly separates masters and non-masters since non-masters (in many cases) will not be given the same (or any) opportunity to demonstrate mastery (or lack thereof) in practice (Kane, 1994). Others have made the case that an examination and corresponding standard would not be needed to properly classify individuals into mastery states if such an external measure existed (Kane, 1994). Given that we cannot validate that a cut score is “right,” the burden of support for a performance standard is not as simple as comparison with another “correct” measure.

Kane (1994) explored validating performance standards as an evaluation of interpretive arguments. Interpretive arguments are those arguments and corresponding underlying inferences that link an examinee’s score to the conclusions made about their competence. He added that an interpretive argument for an appropriate passing standard is based upon two assumptions: “a descriptive assumption, claiming that a passing score corresponds to the performance standard, and a policy assumption, claiming that the performance standard is appropriate” (1994, p. 437). He recommended relying heavily on procedural evidence to support these assumptions. Thus, each step in developing a standard (selection of participants, training, data collection procedures, etc.) should be evaluated, and evidence that procedures were implemented correctly can largely contribute to the validity argument of the standard.
Norcini and Shea (1997) also advocated for standards validation by evaluating the standard setting process components. “To ensure the credibility of the standard, it is important to involve the proper experts and to use a method that (a) produces absolute standards, (b) is based on informed judgment, (c) demonstrates due diligence, and (d) is supported by research” (p. 55). They also noted that evaluating how “realistic” the standard is (a concept parallel to Kane’s (1994) “policy assumption”) is important. They recommended external validation techniques, such as comparison of the standard with other tests that measure similar constructs, as well as the perceived acceptability of the standard by stakeholders.

Drawing from Kane (1994) and AERA, APA, & NCME (1999), Pitoniak (2003) encouraged evaluation of procedural evidence, internal evidence, and external evidence. Procedural evidence, noted earlier, is evidence that appropriate procedural steps were followed. This includes evaluating whether the standard setting purpose and processes were adequately explained, whether participants properly selected and trained, and whether the participants have confidence in the process and results (Sireci, Pitoniak, Meara, & Hambleton, 2000). Internal evidence is a measure of standard consistency and may include measures of inter-rater and intra-rater reliability when making ratings. External evidence consists of comparing standard setting outcomes with expected outcomes and evaluating the perceived reasonableness of the cut scores (Sireci et al., 2000).

C. Existing Standard Setting Method Features and Classification

Many standard setting methods share similar features. Cizek (2012a) points out that all methods incorporate judgments about persons and items. Many setting models include two parts, defining minimal competence and translating that definition to a cut score (quantification). These
will be referred to as “MCE” models of standard setting. Hambleton, Pitoniak, and Copella (2012) list nine standard setting steps common to many currently used methods:

1. Select a standard-setting method and prepare for the first meeting of the panel
2. Chose a large panel that is representative of the stakeholders, and a design for the study.
3. Prepare descriptions of the performance categories.
4. Train panelists to use the method.
5. Collect ratings.
6. Provide panelists with feedback on their ratings and facilitate a discussion.
7. Compile panelist ratings again and obtain performance standards.
8. Conduct an evaluation of the standard-setting process and recommend performance standards.

1. Method classification

Standard setting methods are often sub-categorized into examinee-centered methods, where judgments are made about examinees, and examination-centered methods, where judgments are made about examination items (Jaeger, 1989; Kane, 1994). Because examination-centered methods are most commonly used in licensure and certification, I will explore only the evolution of the commonly used examination-centered methods.

Standard setting methods are typically classified by the type of ratings provided. For example, an Angoff standard setting method and the Nedelsky method may be identical in panel selection and training, but the ratings (i.e., percentage of MCEs that would answer an item correctly versus distractors an MCE could identify as incorrect) distinguish the two. Hambleton, Jaeger, Plake, and Mills (2000) (as cited by Pitoniak, 2003) described a more comprehensive classification. The dimensions of standard setting classification that they outlined included: “(1)
focus of panelists’ judgments; (2) panelists’ judgmental task; (3) judgmental process; (4) composition and size of panel; (5) validation of resulting passing standards; and (6) nature of the assessment” (Pitoniak, 2003, p. 6). Within any method, variations in one of these dimensions (panel size, for example) can lead to different quantifications of the passing standard.

D. A History of Standard Setting Methodology

1. Nedelsky

Arguably the first method for establishing an absolute standard was the Nedelsky method, which was proposed in 1954. This method requires panelists to conceptualize a borderline candidate or MCE, referred to as the F-D student, which is a student who is on the borderline between passing and failing (Nedelsky, 1954). Panelists make item-level judgments about how they think an MCE would answer and, specifically, which options in an MCI the F-D student would be able to eliminate as being incorrect. Research found that this method may not be ideal, as MCEs likely do not randomly guess among the options they are unable to identify as clearly wrong (Livingston & Zieky, 1982; Melican, Mills & Plake, 1989). Additionally, constraining the number of possible probability values for each item is limiting (Brennan & Lockwood, 1980) and may result in standards that are too low (Shepard, 1980).

2. Angoff

William Angoff (1971) introduced the Angoff method, which requires standard setting panelists to first conceptualize a borderline candidate or MCE. Then, the panelists estimate whether an MCE (or a group of MCEs) would answer each exam item correctly by indicating “yes” or “no” for the item. Each item’s ratings are averaged, and then those item averages are averaged to obtain a recommended passing percentage correct. A variation of this method instead tasks panelists with estimating the percentage (0%-100%) of MCEs who would answer each item
correctly. This variation is often referred to as the modified-Angoff method; however, this term is not used consistently in the literature to mean estimating 0-100%. Although there is “no consensus on the definition for the modified-Angoff process” (Reckase, 2000, p. 3), most variations share a general framework. Plake and Cizek (2012) define the “common steps in a traditional Angoff standard setting procedure with MCQs” (p. 186) as: (a) orientation, (b) round 1 ratings, (c) feedback, (d) round 2 ratings, and (e) additional rounds of ratings. The first step involves orienting panelists to the meeting purpose, tasks, and materials, reviewing or developing the definition of minimal competence, and reviewing items and discussing how they would rate them. Next, in round 1, panelists’ ratings for all examination items are recorded. The feedback round includes normative and reality feedback. Additionally, the facilitator may provide impact information to panelists, such as the resulting pass rate if the standard was set using round 1 data. Next, round 2 ratings take place where panelists are instructed to adjust their round 1 ratings based on feedback and discussion. Additional rounds of rating may occur until inter-rater agreement is considered sufficient. Plake (2008) concurs that “most standard setting practitioners support the need for at least two rounds [of ratings], with feedback provided between rounds” (p. 7). Of course, significant variations in this “common” application of the Angoff method are plentiful (Brandon, 2004; Hambleton & Pitoniak, 2006; Norcini, Shea & Kanya, 1988).

3. **Ebel**

Ebel (1972) developed a method in which panelists make judgments about item relevance (essential, important, acceptable, and questionable) and item difficulty (easy, medium, and hard). Based upon these judgments. Items are placed in one of twelve categories (e.g., acceptable/easy, questionable/hard, etc.) The panelists then provide judgments of how they believe MCEs will
perform on items in each of the categories, expressed as a percentage correct. These percent
judgments are weighted based upon the number of items in that category, summed, and then an
average is obtained across all panelists. Shepard (1984) identified the potential cognitive
complexity of separating out items in terms of relevance and difficulty. Cizek (1996) noted that
this method may place undue burden on panelists, given that empirical item difficulty
information is often made available and does not need to be estimated by panelists.

4. **Jaeger**

Jaeger (1978; 1982; 1989) proposed a method that eliminated the need to
conceptualize an MCE. He noted the other methods popular at the time required judgments about
an abstract, artificial MCE group (1982). Instead, he recommended that panelists answer “yes” or
“no” to the question, “Should the group (in question) be able to answer this item correctly?”
(1982). This approach also utilized iterations of ratings and provision of feedback between
rounds. Jaeger and Keller-McNulty (1991) noted that this approach may produce unreasonably
high standards.

5. **Compromise methods**

In the mid-1980’s, several “compromise methods” were developed in an attempt
to methodically combine item performance judgment information and judgments about
reasonable pass and failure rates. While often used in conjunction with an absolute method (e.g.,
Angoff), the two methods described in detail next could be used as stand-alone standard setting
methods.

The Hofstee (1983) method requires panelists to make judgments regarding the minimum
and maximum required passing score on an examination as well as the minimum and maximum
acceptable failure rate. The panelists’ average ratings are calculated and the resulting values
plotted as \((x,y)\) coordinates and connected with a straight line. This line is plotted against a curve of the candidate test-score distribution with the \(x\)-axis value at the intersection indicating the recommended cut score.

The Beuk (1984) method requires panelists to make two judgments. First, panelists indicate the percentage of items that an MCE candidate should be able to answer correctly. Second, the panelists indicate the expected pass rate for the entire examinee population. Like the Hofstee (1983) approach, the panelists’ ratings are calculated and averaged and a line is plotted against a curve of all examinee total test scores. The \(y\)-axis coordinate at the intersection of the lines is the resulting cut score.

6. **Bookmark method**

In the late 1990s and early 2000s, the bookmark method grew in popularity. It was developed in part to reduce the “cognitive complexity perceived to be inherent in the… ubiquitous Angoff method” (Peterson, Schulz, & Engelhard, 2011). With the bookmark method, panelists are presented with an ordered item booklet where examination items are ordered from least to most difficult (usually as estimated by an item response theory (IRT) model). The panelists then “place a bookmark where the items preceding it represent content that the borderline examinee should, with a specified probability, be able to answer correctly” (Hambleton & Pitoniak, 2006, p. 438). Research varies in the recommendations for the specified response probability (RP) to be used when making this judgment; however, values of .50 and .67 are often used (Hambleton & Pitoniak, 2006; Wang, 2003). Hein and Skaggs (2009) found that panelists participating in bookmark standard setting perceived the item difficulty as different than their order in the ordered item booklets (called item disordinality), thus they had trouble determining where to place a bookmark.
7. **Item mapping**

Wang (2003) also noted some of the shortcomings of the Angoff or modified-Angoff standard setting approach, and presented an alternative procedure that is similar to the bookmark method. In the item mapping\(^2\) procedure she proposed, items are calibrated using a Rasch measurement model (Rasch, 1960). As opposed to the traditional bookmark method, which presents items one by one in an ordered item booklet, a histogram (item map) is presented to the panelists. Items with similar difficulty values are grouped into a single column in the histogram. The group reviews items one at a time, starting with an easy item (from the left-most histogram column), moving through the item map to the right. After reviewing an item, the judges are asked whether a typical MCE has at least a 50% chance of answering that item correctly. The process is repeated until the judges reach agreement on what item (or items) located on the map have exactly a 50% chance of being answered correctly by an MCE. Fifty was chosen as the RP for this method because, when the Rasch model is used to calibrate items, item information is maximized at a probability value of .50. The median RIDP of the items selected are transformed into a cut score (Wang, 2003; 2009).

The Rasch model utilizes a different approach than classical test theory to quantify both an examinee’s performance on an examination and item difficulty. The approach ultimately places examinees and items on a common scale, the units of which are called logits. The logit scale maintains the same interval over the continuum, resulting in the property of additivity.

\(^2\) The term “item mapping” is sometimes referred to in the literature as a category of standard setting methods including the bookmark method. For the purposes of this study, “item mapping” will refer to Wang’s 2003 definition.
The dichotomous Rasch model predicts the probability of success for examinee \( n \) on item \( i \):

\[
P(X_{ni} = 1|\theta_n, \beta_i) = \frac{e^{(\theta_n - \beta_i)}}{1 + e^{(\theta_n - \beta_i)}}
\]

Where
\[
\theta_n = \text{ability for examinee } n
\]
\[
\beta_i = \text{difficulty of item } i
\]

Many standard setting methods calculate person and item performance using classical test theory, where information on examinees and items can never be completely separated. In classical test theory, a \( p \)-value (item difficulty) is quantified by the percentage of individuals that answered an item correctly. Although the \( p \)-value measure is often interpreted as a characteristic of an item, it is only characteristic of that item as administered to that set of examinees. Unlike classical test theory, the Rasch measurement model can parse out item difficulty and examinee ability without the caveat that the values are specific to a particular examination with a particular population (Smith, 2004).

E. **Current Standard Setting Practices**

In determining an appropriate standard setting method for a high-stakes MCI examination, psychometricians must balance the considerable amount of literature on methods and their variations, the practicality of the time and resources that can be allocated to standard setting, as well as the *Standards for Psychological and Educational Testing* (hereafter: *Standards*) (AERA, APA & NCME, 2014). The *Standards* clearly promote absolute standards but do not directly make a statement about certain methods, normative data, iterations, or compromise methods.
The Angoff method is arguably the most commonly used standard setting method, and has been asserted as such throughout the last three decades (Mills & Melican, 1988; Plake & Cizek, 2012; Stone, Koskey & Sondergeld, 2011). The widespread use of the Angoff method is not itself a validity argument to establish a passing standard; however, this fact indicates its effectiveness and practicality in implementation. According to Berk (1986), "The Angoff method appears to offer the best balance between technical adequacy and practicability" (p.147), suggesting this method to be the most favored.

The bookmark method has also grown in popularity since its introduction in the late 1990s. It has gained popularity mostly in state educational assessments, including being used for the National Assessment for Educational Progress (NAEP) standard setting (Lewis, Mitzel, Mercado & Schulz, 2012). Karatonis and Sireci (2006) noted that many of the reports of the bookmark method being used in practice came from conference presentations, and more peer-reviewed research is warranted.

1. **Empirical performance data, feedback, and iterations**

   Over the past several decades, standard setting research has begun to address the type and amount of information and training that would be beneficial to panelists. The provision of item performance data, feedback to panelists, and number of rounds of ratings were explored in many studies (Brandon, 2004).

   Researchers investigated what occurs when empirical item performance data, usually item difficulty and/or discrimination, is provided to panelists during standard setting. This data provides information about how the population of examinees performed on each item (e.g., 90% of examinees answered an item correctly). Research has shown that providing this information to
panelists serves as a good “reality check,” which in turn prevents unrealistic standards (Norcini & Shea, 1997; Norcini, Shea & Kanya, 1988; Shepard, 1980).

Additionally, research supports providing feedback to panelists after an initial rating task. Then, a second round of ratings allows panelists to incorporate that feedback. Reckase and Chen (2012) explored several different types of feedback that can be given to panelists after an initial rating task. The first type is data provided to panelists to inform them of how they rated items in relation to other panelists (e.g., a rank-ordered list of average ratings by panelist). The second type is a presentation and discussion about the items where panelists have the greatest disagreement. These data and resulting discussions allow panelists to reconsider their judgment-making process. A third type of data is impact data. This feedback is unique in that it incorporates information about aggregate examinee performance (e.g., test scores). For a pass/fail standard, an impact analysis typically lists the percentage of examinees who would pass if the standard from the first round of ratings was implemented. The purpose of this feedback is to allow raters to assess the standard’s reasonableness. Research shows the use of feedback and iterations improves inter-rater agreement (Busch & Jaeger, 1990; Hambleton, 2001; Plake, Melican & Mills, 1991; Reckase & Chen, 2012). Panelists often report higher task understanding and confidence in their rating ability using this procedure (Plake & Hambleton, 2000; Skorupski & Hambleton, 2005). Brandon’s (2004) literature review found that “judges’ discussions and reviews of empirical information shows that this standard-setting step more often than not affects cut scores, enhances the reliability of item estimates, and reduces their variation” (p. 79). As previously mentioned, the Standards do not advocate for a particular standard setting method, but they do state that:
…feedback on accuracy, the experience of actually taking a form of a test, feedback on the pass rates entailed by provisional proficiency standards, and other forms of information may be beneficial in helping participants to reach sound and principled decisions. (AERA, APA & NCME, 2014, p. 108)

2. **Adjustments**

Another common question in standard setting research is whether making adjustments to cut scores derived from the commonly used models is appropriate. It is not uncommon for a standard setting panel to arrive at a cut score, which is subsequently presented to a panel of policymakers for adoption. Geisinger (1991) and Geisinger and McCormick (2010) explored several reasons why post-hoc adjustments to cut scores might be necessary, such as anomalies occurring in the rating process. Mills and Melican (1988) suggested allowing policymakers to adopt a cut score adjustment relative to the examination’s standard error of measurement (SEM). Selecting a cut score plus one (or more) SEMs would result in fewer false positives (individuals achieve a passing score when they have not actually mastered the content). Policymakers would then have the ability to weigh the relative costs of misclassification errors (Geisinger & McCormick, 2010). Additionally, the use of adjustments to account for unreliability in judgment (standard error of judgment (SEJ)) is often provided to panelists when making final cut score decisions (Geisinger & McCormick, 2010).

**F. Criticisms of the MCE Methods**

Despite many positive treatments of the MCE methods in standard setting literature, criticisms of these methods and their variations exist and require attention.

All the aforementioned trends in methodology (provision of item performance data, feedback, inclusion of iterations, and adjustments) were developed to address the same problem:
low rating reliability. Angoff (1988) noted the reliability in standard setting ratings issue “urgently needs attention” (p. 219). Reacting to the same unreliability state, Cizek (1993) noted, “…there still exists a subtle, discomforting angst among measurement professionals [regarding the way we currently conduct standard setting]” (p. 95). While some degree of refining and re-training panelists throughout the standard setting task is certainly plausible, providing them with a multitude of data that may sway their ratings and then adjust the resulting cut scores should cause serious pause among measurement professionals. Further, increased panelist confidence in the procedure and increased reliability of ratings are often treated in the literature as desirable outcomes or as contributing to the validity argument of the passing standard; however, these findings may not contribute to validity argument.

1. **Inter-rater reliability**

As briefly noted earlier, increased rater agreement may not be a desirable outcome. Such agreement could be the result of two situations. Either the raters are converging on the “true” quantification of the passing standard or the raters are biased by provided data or others’ opinions and are converging their ratings to align with the mean. In other words, the reliability and item rating stability are not a result of a common minimal competence conception among raters. Rather, they are a result of external information overuse in developing the ratings. Clauser et al. (2009) noted that increased agreement of panelists is not necessarily indicative of accuracy in ratings. Brennan and Lockwood (1980) noted the validity arguments made in support of some methods rely too heavily on how much panelists agree with one another. Stone et al. (2011) noted that iterations aimed at consensus building defeat the purpose of the varied opinions of a diverse panel.
2. **Panelist confidence**

   Increased panelist confidence in ratings likely supports the passing standard interpretation validity; however, it is possible that the confidence is a result of the consensus-building exercises. When individual differences are minimized and an individual’s ratings match the group’s ratings, this may instill some confidence that the panelists agree. However, this agreement may not be the result of the application of the true minimal competence definition.

3. **Empirical performance data, feedback, and iterations**

   In both providing empirical item data and feedback based upon examinee performance (impact analysis), norm-referenced information is provided to panelists to inform their judgments. The same reasons why norm-referenced standards are not appropriate for criterion-referenced examinations support the argument that norm-referenced data should not be a part of absolute standard setting. Hambleton and Pitoniak (2006) note that some policymakers believe standards should be set without the knowledge of potential consequences.

   Lastly, the frequent use of compromise methods or other adjustments suggests that cut scores derived from the judgmental methodology alone result in unrealistic or unreasonable standards. Stone (2001) concurs, "[t]he use of statistics to correct a criterion standard suggests something is wrong with the standard and defiles its meaning in many ways while making a mockery of the scientific process" (p. 197).

4. **Minimal competence**

   The previous critiques of the Angoff method focus on the methodology trends aimed at consensus-building and adjusting scores to develop “reasonable” standards. However, an even more fundamental assumption in the method requires attention: the concept of minimal competence.
An unanswered question about all MCE methods is whether making judgments about minimal competence is cognitively possible. To determine whether the concept of minimal competence is being applied consistently among raters, studies have evaluated the reliability of ratings within raters, within panels and whether those ratings are the same on different occasions (Norcini & Shea, 1992; Plake & Impara, 2001).

Norcini and Shea (1992) found Angoff estimates remain very similar over several rating occasions, which provides evidence that panelists are applying a consistent standard (i.e., have a common understanding of the MCE). Plake and Impara’s 2001 study had similar findings. Conversely, Glass (1978) argued that a conception of minimal competence is not possible. “The attempt to base scores on the concept of minimal competence fails for two reasons: (1) it has virtually no foundation in psychology; (2) when its arbitrariness is granted but judges attempt nonetheless to specify minimal competence, they disagree wildly” (p. 251). Brandon (2004) concluded from a review of studies that panelists were unable to apply judgmental rules consistently. Impara and Plake (1998) found panelists who set a teaching examination standard were more accurate when estimating the total examinee group’s performance than the “borderline” group (MCEs). Plake, Impara, and Irwin (2000) concluded that panelists were internally consistent in their ratings of hard and moderately hard items, but they were less consistent for easy items. Skorupski (2012) noted the Angoff method’s reliance on a definition of minimal competence has been “widely used and widely criticized as being very cognitively challenging for panelists” (Skorupski, 2012, p. 139).

Stone (1996) argued that panelists are experts in content; not psychometrics. Asking panelists to make speculations about an MCE’s performance on an item places them in a “fortune-teller” or psychometrician role. Speculating an examinee-item interaction is not
consistent with measuring the construct of interest. “The original notion of a construct as an idealized criterion is lost. Here in its place is the examinee-item interaction as the criterion. We can only vaguely hope that the real construct that we seek to measure resembles this interaction” (Stone, 1996, p. 6).

McGinty (2005) found that panelists had trouble making a distinction between the percentage of MCEs who would answer an item correctly (prediction) versus the percentage who should answer correctly (value judgment). She noted that the research on standard setting using qualitative measures and drawing on research in cognitive psychology is lacking, and much is still unknown about the thought processes of standard setting panelists.

Given these observations, standard setting facilitators who apply the MCE method must take great care in evaluating the understanding of the panelists and determining whether they are able to reliably apply a definition of minimal competence.

G. Method Comparisons

Arguably the two most popular methods being used today in educational and licensure and certification testing are variations of the Angoff method and the bookmark method. Given the merits of each, a handful of studies have compared these two methods for setting (a) standard(s) on a single assessment.

Buckendahl et al. (2002) compared the Angoff and bookmark standard setting methods by using two different panels (once using each method) to set a standard on a Grade 7 mathematics assessment. Both methods utilized two rounds of ratings. The recommended cut scores for each method were similar, but the bookmark method’s cut score had a slightly lower standard deviation than the Angoff method, suggesting more panelist agreement. The authors noted some limitations of their study, including the application context (education, lower-stakes
assessment), which may not generalize to other fields. In a study comparing the Angoff and bookmark standard setting methods for the NAEP, Peterson et al. (2011) found that bookmark-based methods were comparable to the Angoff method in reliability and resulting cut scores. Furthermore, the bookmark method took less time to conduct and was determined to be a more efficient method for future standard setting studies. In a literature review of bookmark method standard setting procedures, Karantonis and Sireci (2006) noted that the bookmark method was often rated highly in terms of panelist confidence, but it could also result in standards that were determined to be too low. Wang (2003) found that the cut scores set with the item mapping method were lower than those set with the Angoff method. Her study included four standard setting comparisons, and the item mapping cut score was lower than the Angoff in each comparison by between 9% and 19% of the total items on the examination. She also noted that “[j]udges agreed that the item-mapping method set more realistic cut scores than the Angoff method” (p. 248).

The comparisons of the Angoff and bookmark methods are challenging to interpret, partly because each study implements the methods slightly differently. In general, the bookmark method was found to result in comparable or lower standards than the Angoff method. In several cases, the bookmark standard was considered to be more reasonable in terms of the resulting pass rates.

H. **Summary of Literature**

This study compares the item mapping method to the Angoff method with application to a recertification examination. There are several reasons why the item mapping method (as presented by Wang, 2003) was chosen over the traditional bookmark method: first, the item mapping method provides panelists with a histogram as opposed to an ordered item booklet. This
provides a “global picture of all items and their estimated difficulties… which serves to guide and simplify the judges’ process of decision making during the cut score study” (Wang, 2003, p. 232). Secondly, the use of the Rasch model to calibrate test items and examinees allows a straightforward interpretation of the task and judgment to be made. When examinee ability equals item difficulty, the likelihood of answering correctly is 50%. “By utilizing the Rasch model’s distinct relationship between candidate ability and item difficulty, the item-mapping method enables judges to determine the passing score at the point where item difficulty equals the MCC’s ability level” (Wang, 2003, p. 232). It is evident that the item mapping method affords some advantages over the Angoff and other MCE standard setting models, in theory. However, the Angoff method has been established as a credible standard setting method in licensure and certification assessments, and certain operational aspects of the item mapping method require more evaluation. Comparing two standard setting methods is known to add external validity evidence to the interpretation of the resulting cut scores (Pitoniak, 2003; Hambleton, Pitoniak & Coppella, 2012). Given the limited number of studies comparing the item mapping method and the Angoff method, additional research is needed to evaluate the merits of each.

This research adds valuable information to the field of high-stakes licensure and certification testing on the relative benefits of these two standard setting methods. Additionally, the findings may influence standard setting in examination contexts outside of licensure and certification.
I. METHOD

A. Research Question

This research intends to answer the question: “Which standard setting method (item mapping or Angoff) results in a more appropriate cut score on this high-stakes MCI examination?”

To answer this question, both standard setting methods must be evaluated with respect to procedural validity, internal validity, and external validity. To establish an argument for procedural validity, I documented that the appropriate procedural steps were followed for both methods. This included selecting a diverse and representative panel, providing complete training prior to the rating tasks, and surveying panelists to determine if they understood the methods and tasks. With respect to internal validity, I determined which method’s use resulted in higher inter-panelist and intra-panelist consistency. With respect to external validity, I evaluated which method yielded a cut score that was perceived as more reasonable by the panelists and was more consistent with expectations about the performance of the population.

B. Examination Properties

The examination used in this study was one form of a medical recertification examination in ophthalmology, which is administered annually to approximately 800-1000 examinees. Individuals who have received initial certification in the field must pass this recertification examination at least once every ten years. The examination administered actually consists of three separate, 50-item examinations or “modules.” All examinees are required to take one 50-item module that covers core knowledge needed by all certificate holders. Examinees may then choose two 50-item modules from a list of ten, which cover knowledge that is specific to a certain area of sub-specialty practice. A unique cut score is developed for each module, and the
examination is scored compensatorily based on raw number of items answered correctly (out of 150). For example, let an examinee’s performance on three modules be 40 items correct, 45 items correct, and 35 items correct on the core module, first sub-specialty module, and second sub-specialty module, respectively. If the cut scores for those three modules are 40, 40, and 39, this examinee would pass the examination, as their total score (120) is greater than the sum of the cut scores for the three modules they took (119). The 50-item core knowledge module that is taken by all examinees was the module used in this study. This module will hereafter be referred to as the “examination.”

The examination in this study consists of 50 four-option single best answer MCIs. All 50 items were scored items; no items appeared on the form as unscored (pre-test) items. All examinees answered the same 50 scored items. Items were scored dichotomously. The form of the examination was administered to 814 examinees in 2016. The range of scores was 27 to 49, \((M = 42.47, SD = 3.36)\). The performance distribution of raw scores achieved is depicted in Figure 1.
Examinees were given one hour to complete the 50-item examination. The examination was administered throughout the month of September of 2016 via computer at testing centers across the United States. The examination was assembled using a content blueprint containing ten subject domains, each of which make up a specific percentage of items on the examination.

Each year, a cut score is established for this examination by use of the Angoff standard setting method (no equating to a previously administered form takes place). After standard setting, panelists arrive at a recommendation for a cut score, then the cut score is reviewed and
approved by a subset of the certifying board’s Board of Directors. This study did not make use of the data (Angoff ratings) obtained in setting an operational standard for this examination. Rather, the study arrived at new cut scores that were not used in practice. In 2016, 92.1% of examinees passed this examination module. The pass rate for the 814 examinees on the full 150-item examination was 95.9%.

C. **Study Design**

This study compared two standard setting methods and their resulting cut scores using examination content and performance data from the examination administered to 814 examinees in 2016. All panelists rated all fitting examination items using both standard setting methods: the Angoff method and the item mapping method. Half of the panelists (group 1) rated items using the Angoff method first and the item mapping method second. The other half (group 2) rated items using the item mapping method first and the Angoff method second. This study design is recommended by Wang (2003): “To provide a better design in comparing the two methods, two separate groups of judges [panelists], each using one presentation order to set a cut score for the same exam should be considered for future research” (p. 246).

All panelists (both group 1 and group 2) participated in the same initial training activities, but engaged in method-specific training and ratings tasks separately.

1. **Items**

The items on this examination are meant to measure knowledge of core knowledge needed by certificate holders. An examination blueprint, which consists of ten subject domains, is used to assemble the examination so that content areas are represented appropriately (see Table I). In preparation for this study, all of the items’ performance statistics were calculated using both classical test theory and the Rasch measurement model. All statistics were
calculated based upon the performance of all 814 examinees who attempted this examination in 2016. The reliability of the form estimated with Cronbach’s alpha (KR-20) was $\alpha = .52$.

**TABLE I**

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Number of Items on Form</th>
<th>Target Number of Items per Blueprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataract and Anterior Segment</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Cornea and External Disease</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Glaucoma$^a$</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Neuro-Ophthalmology and Orbit</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Oculoplastics and Orbit</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Ophthalmic Pathology and Oncology</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Pediatric Ophthalmology and Strabismus</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Refractive Management and Optics</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Retina and Vitreous</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Uveitis</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

$^a$ One item from the Glaucoma section was removed from the standard setting exercises due to misfit; however, that item is included in this table.

First, classical test theory values of performance of the entire examinee population, including $p$-values (difficulty) and biserial correlation values (discrimination), were calculated. A $p$-value ($p$) is the percentage of individuals who answered an item correctly.
The biserial correlation \( r_b \) is:

\[
r_b = (Y_1 - Y_0) \times \left( \frac{pq}{\sigma_Y} \right)
\]

Where:

- \( Y_1 \) = mean score on examination for examinees who answered item correctly
- \( Y_0 \) = mean score on examination for examinees who answered item incorrectly
- \( p \) = percentage of examinees who answered item correctly
- \( q \) = percentage of examinees who answered item incorrectly
- \( Y \) = height of the standardized normal distribution at the point \( z \), where \( P(z' < z) = q \) and \( P(z' > z) = p \)
- \( \sigma_Y \) = the population standard deviation

For some examinations, the point biserial value is reported as the measure of item discrimination as opposed to biserial correlation. While both statistics measure the correlation between two variables, the biserial correlation is recommended when one of the variables is artificially dichotomous (e.g., overall test performance as pass/fail) (Brown, 2001).

RIDP values (for an item \( i \), difficulty (RIDP) for an item is denoted \( \beta_i \) in the aforementioned dichotomous Rasch model) were calculated using WINSTEPS version 4.0.0 (Linacre, 2016). Using the Rasch model, item discrimination is assumed to be constant, which is why there is no Rasch-equivalent value that can be calculated with regard to item discrimination. Item fit statistics were also calculated. These include the standardized weighted total fit and standardized unweighted total fit statistics, which assess how well the data fits the Rasch model. Specifically, these statistics show the difference between the observed and expected frequency of examinees answering an item correctly (Wright & Stone, 1979). Any item with either standardized fit statistic greater than 2 or less than negative 2 was excluded from use in ratings of
both methods so that comparisons of cut scores are based upon evaluating the same group of items (Wang, 2003). One item (item 6460) was removed from the standard setting ratings due to misfit (infit ZSTD = 2.5, outfit ZSTD = 2), but it is still included in the overall examination performance tables because the item was included in calculating final examination scores. Overall examination item and person performance data are reported in Tables II and III.

**TABLE II**

<table>
<thead>
<tr>
<th>Examination Form Item Performance, $N = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean $p$-value (SD)</td>
</tr>
<tr>
<td>Mean biserial correlation value (SD)</td>
</tr>
<tr>
<td>Mean RIDP (SD)</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th>Examination Form Person Performance, $N = 814$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean $p$-value (SD)</td>
</tr>
<tr>
<td>Mean Rasch person ability estimate (SD)</td>
</tr>
</tbody>
</table>
After the initial item analysis was completed, subject matter experts reviewed and validated keys for items flagged as having statistical performance outside defined parameters. Items were flagged for key validation if they had a \( p \)-value less than or equal to 0.6 or a biserial correlation value less than 0. Four items on this exam form (items 6460, 6474, 6595, and 5364) were flagged for having \( p \)-values less than or equal to 0.6 (\( p=0.60, 0.38, 0.60, \) and 0.57, respectively), and no items were flagged for having a negative biserial correlation. All items were reviewed by a group of subject matter experts, and no items were determined to have more than one key or a different key than originally documented.

**Participants (panelists)**

Sixteen total panelists were recruited to participate in the study, with eight in each presentation order condition (group). Although most research acknowledges that more panelists are usually preferable, five to ten panelists have been found to be acceptable (Brennan & Lockwood, 1980; Norcini, Shea, & Grosso, 1991). To be eligible to participate, panelists had to have completed the requirements for initial certification, which included successful completion of an accredited training program, passing one MCI examination and passing one oral examination. A pool of individuals who expressed interest in volunteering for the certification board by writing examination items were asked via e-mail if they were interested in participating in an optional research study. The e-mail was sent to approximately 80 item writers approximately one month prior to the beginning of the study. Twenty individuals expressed interest, and sixteen were selected based upon their availability to attend one of two item mapping webinars scheduled for evenings in July and August 2017. All panelists who participated in the study had experience writing items for one or more of the certifying board’s examinations.
Research on the ideal composition of a standard setting panel is varied. Many studies support using only panelists who are subject matter experts in the field of the examination (Brandon, 2004; Hambleton & Pitoniak, 2006; Plake, 2008). Per Norcini, Shea, and Kanya (1998), in a field with considerable specialization, use of panelists who are sub-specialists (as opposed to general or comprehensive practitioners) is acceptable, as it does not affect their estimates of expected performance using the Angoff method. Additionally, research supports the use of panelists who are familiar with the knowledge, skills, and abilities of the target population of the examination (Plake, 2008). Given these guidelines, the above criteria for panelist eligibility and recruitment in this study is appropriate. However, some limitations are noted in the discussion chapter related to the recruitment process.

D. **Standard Setting Procedures**

In the next sections, the procedures for training and conducting the Angoff standard setting and item mapping standard setting activities are outlined.

1. **Schedule**

   The standard setting activities occurred in two phases: 1) initial training, and 2) rating tasks. Initial training occurred remotely with all panelists in both groups. Then, the Angoff and item mapping rating tasks (and training specific to those rating tasks) occurred remotely with each group individually. Within two weeks of completion of the initial training, group 1 was trained on the Angoff methodology and conducted Angoff ratings. Less than one week following initial training, group 2 was trained on the item mapping methodology and participated in the item mapping ratings. Within two weeks of completion of these tasks, the groups participated in ratings using the other method (group 1: item mapping, group 2: Angoff). Lastly, a conference
call discussion was held among all of the panelists. This call occurred one week after all of the ratings were compiled, which was approximately four weeks after the initial training meeting.

Figure 2. Standard setting activities schedule.

2. **Initial training**

The initial training session occurred via an online webinar, which was attended by 11 panelists. The remaining five panelists who were unable to attend watched a recording of the presentation and group discussion. This training introduced the purpose of standard setting and engaged the panelists in developing a group consensus on the definition of the borderline
examinee (Appendix A). The term “borderline” is synonymous with the MCE but is the term preferred by the testing organization.

First, panelists were presented with a definition of standard setting and purpose of standard setting. Next, the group was provided with information on the examination format, as well as the performance of examinees. The performance of examinees was depicted as a histogram of the examinees’ performance (raw scores on the examination). The group then engaged in a discussion regarding the definition of a borderline examinee. This discussion involved brainstorming and creating a list of the knowledge, skills, and abilities that must be possessed by examinees in order to pass this examination, as well as those knowledge, skills, and abilities that are desirable but not essential to possess. This is consistent with the recommendation by Hambleton, Pitoniak, and Coppella (2012): the approach for developing a definition of minimal competence should always highlight what needs to be known to meet the performance standard and be documented in writing.

Lastly, the group was trained on interpreting item performance data and reviewed the content of several examination items used on previous examination forms. Five items (stem and options) were shown to the panelists one at a time. For each item, after presentation of the item content and some discussion about the perceived difficulty of the item by the panelists, classical test theory performance data ($p$-values and biserial correlation values) were shown. The conditional $p$-values for these items were also presented. Conditional $p$-values in this context refer to the percentage of borderline examinees who answered the item correctly. Borderline examinees were considered individuals who achieved a raw score equal to or one item greater than the operational cut score of the examination. The panelists were encouraged to discuss what factors make an item more complex, how the borderline examinee typically performs on an item
as compared to the average examinee, and how they would apply the borderline examinee definition to these items. This step, often referred to as practice in making judgments using the definition of minimal competence, is recommended to occur at the end of the initial training (Loomis, 2012).

Following initial training, panelists completed an online survey (see Appendix B). The purpose of this survey was to ensure that panelists understood the training and were prepared to move on to applying their definition of the borderline examinee to examination items (the rating task). Panelists were asked if the training in the standard setting purpose and methods was clear, whether they had a good understanding of the borderline examinee and whether they felt comfortable applying the definition of the borderline on examination items.

3. **Method 1 (Angoff)**

The Angoff method training began with a review of the definition of the borderline examinee. Panelists were then shown two sample items with p-values and biserial correlation values for those items, and asked to think about the likelihood that a borderline examinee would answer those items correctly. This review was limited because it was primarily a review of the information received in the initial training. Immediately following this Angoff-specific training, panelists were given a two-week window to login to a secure online interface in order to review all examination items and input their Angoff ratings. They rated the 49 examination items that fit the Rasch model in the aforementioned item analysis.

Panelists independently rated the examination items using the modified-Angoff method. For each item, they were asked to answer the question, “What percentage of borderline examinees would answer this item correctly?” Panelists were asked to make judgments about the likelihood that a borderline examinee would answer an item correctly on a scale of 0-100. This
version of the Angoff method was chosen over the yes/no version because it had been used in the past for standard setting on the examination. Also, it allowed panelists to contrast the process of developing 0-100 ratings (Angoff) with making yes/no ratings (item mapping). No psychometric performance data was provided during this first round of ratings. Immediately after making their rating on an item, performance data (the \( p \)-value and biserial correlation) were shown for the item. Panelists were then able to adjust their initial rating based upon this additional information.

The second Angoff rating provided by each panelist was used in calculating the final cut score for the Angoff method. Specifically, the average rating for each panelist in round 2 was calculated, and the recommended cut score was the grand mean of panelists’ mean ratings.

Research supports providing panelists with item performance data (Norcini, Shea, & Kanya, 1988); however, it is possible that having item performance data may compete with the panelists’ ability to apply the borderline examinee definition (Reckase, 2001). Because two rating rounds were used for this method, performance data was provided in the second round only, and panelists completed the task given both conditions of “no performance data” and “performance data provided.” Existing research does not clearly advocate providing empirical item difficulty data during particular iterations. However, it does recommend its provision at some point in the rating exercise (Plake & Impara, 2001; Plake, 2008; Clauser et al., 2009), with some research supporting that it should be considered “feedback” and provided after some ratings have been made (Hambleton & Pitoniak, 2006). Providing this data in the second round only allows panelists a “reality check” on their original judgments. Panelists were able to see their ratings from round 1 when making their round 2 ratings.

Some studies support the use of at least two rounds of ratings with some type of normative feedback provided between rounds (Plake, 2008; Reckase & Chen, 2012; Zieky,
This feedback can include impact analyses (pass rates) or discussion of items with large variance in ratings. However, other research indicates that the provision of normative information between rounds is questionable, given that the panelists’ expertise and judgment about the examination content alone is what the standard intends to reflect (Skorupski, 2012). McGinty (2005) also notes that discussion between iterations can result in pressure from more vocal panel members, which could sway the judgments of the other panelists. For this study, introduction of norm-referenced information for cut score development (i.e., use of the Hofstee standard setting method) is used in a future step in this study; thus, provision of norm-referenced information between rating rounds was not included.

Following completion of the second round of Angoff ratings, panelists completed an online survey (see Appendix C). The survey gathered feedback from panelists on their ability to apply the method appropriately.

4. **Method 2 (Item mapping)**

The item mapping method utilized group work to establish a cut score; thus, a webinar was held for panelists to participate in remotely. Like the beginning of the Angoff method training, item mapping training began with a review of the definition of the borderline examinee. Next, panelists were provided with a brief overview of the item mapping method, including an explanation of the graphical representation (histogram) of item performance. Panelists were told that for this method, items were lined up in order of item difficulty (easiest on the left, hardest on the right). Items within the same column were of very similar levels of difficulty. Panelists were also told “[t]he goal of item mapping is to locate a column of items on the histogram where judges [panelists] can reach consensus that the MCC [borderline examinee] has a .50 chance of answering the items correctly” (Wang, 2003, p. 233). This training as well as
all subsequent item mapping method rating activities occurred during a 90-minute webinar meeting among participants.

After item mapping method training, the histogram or “item map” was presented to panelists (Appendix D). Item difficulties for all items were estimated for all items using the Rasch model, as previously noted. RIDPs (logit values) were then converted into a new scale, with the original logit difficulty value multiplied by 10 and then added to 100. This results in a score scale that ranges from approximately 70 to 130. To develop the columns of the histogram, items were grouped into columns that have scaled difficulty values within a two-point range (e.g., one column of the histogram will include those items with scaled difficulty values of 70-72). Within each column of the histogram, items were color-coded according to content area (specifically, the ten areas of the content blueprint of this examination). All aforementioned steps pertaining to histogram development are consistent with Wang’s 2003 study design. The item map used in the study was created using a table in Microsoft Word. Each item was identified by a unique four or five-digit number. The items in the item map were hyperlinked to the item text on a subsequent page in the document. After each item was reviewed, the map was presented again, and another item was selected for review.

As outlined, individual items were selected from the map one at a time and presented to the group. The item from the left-most column was selected first, and the panelists were asked whether the typical borderline examinee had at least a 50% chance of answering this item correctly. Individuals were asked to privately make note of their answer (“yes” or “no”) and then announce their ratings to the group one at a time. If the group did not reach a consensus (defined as agreement by at least 7 out of the 8 panelists), the $p$-value and biserial correlation value for the item were provided to the group, and the panelists discussed their ratings and then re-rated based
upon this information (Wang, 2003). The final ratings for these items are recorded as “1” for “yes” and “0” for “no” for each panelist. Several items from the same column on the item map were selected, and the process was repeated until the majority of items in that column had ratings shared by the majority of group members. The process was repeated with the next column to the right, which consisted of slightly more difficult items. If a column was reached in which all items were rated and a majority of items were not rated as mostly “yes” or mostly “no,” the panelists were asked whether, for items in that column, “the probability of a correct answer for the MCC [borderline examinee] was .50” (Wang, 2003, p. 243). If the majority of panelists agreed the answer to this question was “yes,” then the item difficulty for the item in the middle of this column, converted from the logit scale value into a raw score, would become the recommended cut score. If the panelists did not agree, the rating process was continued with the next column to the right. Ideally, the response pattern of panelists would shift from mostly “yes” responses to mostly “no” responses as they moved through the columns of items, and the transitional column would be identified. If after rating all columns no transitional column was identified, those columns for which the greatest variance in ratings were present would be re-reviewed and rated. The re-review and re-rating would continue until a majority of panelists agreed on a column that they felt the borderline examinee had a probability of a correct response of 50%.

In the traditional item mapping method, a second, normative step is introduced. After the cut score is established as above, the group is provided with the pass rate corresponding to the cut score developed using the item mapping method. Then, they are asked about the reasonableness of the cut score. This step did not occur as part of the item mapping webinar activity. Rather, it occurred later when panelists were given the opportunity to comment on the reasonableness of both the item mapping and the Angoff cut scores. Additionally, because the
Angoff method did not make use of an impact analysis between rounds, it was determined that for equal comparison of both methods, no information about conditional pass rates for different cut scores would be provided to panelists until all ratings were completed.

Following completion of item mapping ratings, panelists completed an online survey (see Appendix E). The survey gathered feedback from panelists on their ability to apply the method.

5. Evaluation

After the rating tasks were completed by both groups, a fourth, final online survey was administered (Appendix F). This survey consisted of six questions, and allowed for panelists to leave free-text comments. The first four questions related to the pass rates on this examination that panelists felt were acceptable given their knowledge of the examinee population (Hofstee, 1983). After answering these questions, the panelists were presented (within the online survey interface) with the cut scores and corresponding pass rates, which were arrived at using the Angoff method and the item mapping method. Upon reviewing this information, they were asked to indicate their level of agreement with the following: “The recommended cut score(s) using the Angoff method fairly represents the minimal level of performance” and “The recommended cut score(s) using the item mapping method fairly represents the minimal level of performance.”

After the rating tasks and surveys were completed, panelists were asked to participate in a one-hour conference call to discuss the survey results. They were first asked to discuss the Angoff method (and were re-prompted to answer the questions from Appendix C). Then, they were asked to discuss the item mapping method (and were re-prompted to answer the questions from Appendix E). Next, they were presented with the online survey questions and responses from Appendix F one at a time. After each survey question’s response was presented, panelists were asked if they had additional feedback relating to that survey question that they would like to
share. The discussion was transcribed and reported by grouping comments into common themes, which is consistent with survey gathering and reporting techniques used by Skorupski and Hambleton (2005) in evaluating standard setting panelist feedback.

6. **Surveys**

As noted, four brief online surveys were administered throughout the standard setting process. Cizek (2012b) notes: “Administration of ... two evaluations ... [one before making ratings and one after] can provide sufficient information about the standard setting activity. However, when feasible, it is most desirable that several distinct evaluations, administered at carefully identified junctures in the standard setting activity, be conducted” (p. 171).

All survey questions used in this study were adapted from Cizek, (2012b), which he notes are adapted from standard setting studies conducted by Plake, Impara, Cizek, and Sireci (2008). Cizek (2012b) recommends a four or five-point Likert-scale for use on such evaluations. He also notes that a four-point scale may be preferable to a five-point scale (with a fifth, “neutral” category being omitted) as it may force “respondents to make a choice and may more accurately reveal overall perceptions of the participant group” (Cizek, 2012b, p.172). All survey questions used in this study included a four-point Likert scale with the response options “Strongly Disagree,” “Disagree,” “Agree” and “Strongly Agree.” At the conclusion of each survey, panelists also had the option to enter additional comments. The surveys were developed using the secure, web-based interface Qualtrics (www.qualtrics.com).

E. **Analysis Approach**

To analyze the data, both methods and their results were compared in several ways. First, an initial data analysis was used to determine if the groups’ differential ordering of methods
impacted the resulting cut scores. Each was also evaluated with respect to procedural validity, internal validity, and external validity.

1. **Initial data analysis**

After all rating data was received, cut scores for each condition (group 1 Angoff, group 2 Angoff, group 1 item mapping, and group 2 item mapping) were calculated. First, I determined whether the two Angoff cut scores and the two item mapping cut scores were significantly different using a two-sample $z$ test for difference of item difficulty. A difference would indicate that the order of method presentation may have affected the cut scores. If a difference was found, subsequent analyses and results would compare all four cut scores separately.

2. **Procedural validity evaluation**

Procedural validity is established through evidence that appropriate procedural steps were followed. The Angoff and item mapping methods cannot be compared with one another with respect to certain aspects of procedural validity because these procedural steps occurred simultaneously for both methods (i.e., only one panel was selected, and the definition of minimal competence was created once for the study). Procedural validity is established for this study as a whole by documenting the composition of the panel (time since initial certification, subspecialty area of practice, gender, ethnicity, and geographic location). The ideal panel for this study would have one representative from each of the ten sub-specialty practice areas identified by the testing organization as existing in the field and would include an approximately equal distribution of individuals defined as early career (in practice for 20 years or less) and late career (in practice for 21 years or more), due to possible differences in rating behavior by individuals who are more and less advanced in their careers (Brandon, 2004).
Secondly, the panel was asked whether the standard setting purpose and processes were made clear (Appendix B, question 1), whether after defining minimal competence (the borderline examinee) and completing the practice exercise, the definition of the borderline examinee was clear (Appendix B, question 2) and whether the individual panelist felt able to apply the definition of minimal competence to examination items (i.e., rate how borderline examinees would perform (which is required in both methods)) (Appendix B, question 3). These questions were adapted from Cizek (2012b). Success for these three questions was operationalized as at least 75% of individuals indicating “Agree” or “Strongly Agree” for each of these survey questions. The literature on standard setting evaluations has not established a criterion for success on such evaluation questions; thus, the threshold of 75% has been selected subjectively. The frequency of each survey response is reported, as will the median and interquartile ranges (IQR) of the ratings on the score scale (with Strongly Agree = 4 and Strongly Disagree = 1). Because the training for both methods occurred simultaneously and the definition of minimal competence applies to both methods, these survey questions were asked of all panelists once, at the conclusion of initial training, prior to engaging in rating the examination items.

After panelists engaged in the rating tasks for a single method (Angoff or item mapping), they completed another survey. Panelists were asked on a four-point Likert scale how much they agreed with two statements: “I understood how to make my item mapping/Angoff ratings” and “Overall, I am confident that I was able to apply the standard setting method appropriately.” The frequency of each response is reported, as are the median and interquartile ranges of the ratings on the score scale (with Strongly Agree = 4 and Strongly Disagree = 1). The results of these survey questions were compared with the Mann-Whitney U significance test to determine whether panelists felt more confident in their ability to apply one method than the other.
Panelists were given the option to submit additional comments in a free text response box at the end of each survey. Those comments are reported. Common themes from the surveys and the post-rating evaluation conference call are presented and interpreted in context of either supporting or refuting procedural validity for each method.

3. **Internal validity evaluation**

   Internal validity is established through evaluating intra-panelist and inter-panelist consistency.

   Intra-panelist consistency indicates the degree to which individual panelists are able to provide consistent ratings across items. For each panelist, round 2 Angoff ratings and item mapping ratings were correlated with empirical item difficulty values ($p$-values). The average correlation value for each method was compared with Fisher’s $r$- to $z$- transformation, followed by a $z$ test for difference of correlations to determine if one method yielded stronger intra-panelist consistency.

   Inter-panelist consistency was assessed by comparing the standard error of judgment (SEJ) for each method’s ratings. The SEJ indicates the variability of the panelists’ standards, which estimates “the extent to which the performance standards would vary if the study were replicated with many different samples of panelists of a similar size of panel” (Hambleton et al., 2012, p. 62). A lower SEJ indicates higher inter-panelist consistency. The SEJ is calculated as follows:

   $$SEJ = \frac{\sigma}{\sqrt{n}}$$

   Where

   $\sigma = $ standard deviation of panelists’ judgments

   $n = $ number of panelists
4. **External validity evaluation**

External validity is the degree to which the cut score is consistent with expected outcomes (i.e., is reasonable). The reasonableness of the cut scores was evaluated through survey questions (Appendix F, questions 5, 7, and 8) and through discussion in the post-evaluation conference call among panelists. The cut scores were also compared to the Hofstee cut score.

a. **Assessing reasonableness**

The reasonableness of each cut score was assessed by comparing the pass rates determined by the Angoff and item mapping methods to data gathered in the survey questions 5 and 7 in Appendix F, both which ask whether the pass rate arrived at fairly represents the minimal level of competence. The frequency of each response for each survey question is reported, as is the median and interquartile ranges of the ratings on the score scale (with Strongly Agree = 4 and Strongly Disagree = 1). The results of these survey questions were compared with the Kruskal-Wallis H significance test to determine whether panelists felt that any cut score was more or less appropriate than another. Additional comments gathered from open-ended survey comments (Appendix F, questions 6 and 9) and the post-rating evaluation conference call discussion in response to these two survey questions are reported and discussed. Common themes from the survey and the post-rating evaluation conference call are presented and interpreted in context of either supporting or refuting external validity.

In addition, I report the historical operational cut score and corresponding pass rate for this examination (2016 administration) as well as the cut score and corresponding pass rates for its 2014 and 2015 forms. Of note, the 2014 and 2015 forms were not equated or determined to be equivalent in difficulty; however, provision of this information will provide additional background information on the traditional pass rates for this examination.
b. **Hofstee method comparison**

In addition to the above analyses, the Hofstee (1983) compromise method of establishing a standard was also conducted to determine which method(s) cut score(s) is comparable with cut scores arrived at using another method, which adds to the external validity argument (Sireci et al., 2000). The Hofstee method requires panelists to answer four questions about expected performance of the examinee population after reviewing the examination items. These four questions were asked after all rating data was collected. These questions (Appendix F, questions 1-4) asked the panelists for their judgments of the minimum and maximum acceptable passing scores and the minimum and maximum acceptable failure rates on the examination.

To arrive at a cut score using the Hofstee method, the following were calculated from these survey questions:

\[ k_{\text{max}} = \text{highest cutoff score that would be acceptable (mean of panelists’ ratings for Appendix F, question 1)} \]
\[ k_{\text{min}} = \text{lowest cutoff score that would be acceptable (mean of panelists’ ratings for Appendix F, question 2)} \]
\[ f_{\text{max}} = \text{maximum acceptable failure rate (mean of panelists’ ratings for Appendix F, question 3)} \]
\[ f_{\text{min}} = \text{minimum acceptable failure rate (mean of panelists’ ratings for Appendix F, question 4)} \]

Using the points \((f_{\text{min}}, k_{\text{max}})\) and \((f_{\text{max}}, k_{\text{min}})\), a line was plotted on top of an examinee test-score distribution curve, with examinee raw test scores (percentage correct) on the \(x\)-axis and corresponding percentage of examinees who would fail on the \(y\)-axis. The \(x\)-axis value at the intersection of those two lines indicates the recommended cut score. The difference between the
Angoff and Hofstee cut score and the difference between the item mapping and Hofstee cut score are reported. Z tests for difference of item difficulty were used to compare the Hofstee cut score to the Angoff and item mapping cut scores.

F. **Summary of Methodology**

This research intends to answer the question: “Which standard setting method (item mapping or Angoff) will lead to a more appropriate cut score on this high-stakes MCI exam?”

To answer this question, each standard setting method must be evaluated with respect to procedural validity, internal validity, and external validity. To evaluate procedural validity, both methods are evaluated with respect to their use of a diverse and representative panel and the indication of panelists that they understood the rating tasks they were asked to engage in. Secondly, to evaluate internal validity, the methods’ ratings’ inter-rater consistency and intra-rater consistency were calculated and compared. Lastly, external validity was evaluated through surveying the panelists to determine which method’s cut score is perceived as more reasonable. External validity was also evaluated by comparing the methods’ cut scores with the cut score determined by the Hofstee method.
II. RESULTS

This chapter outlines the results of the standard setting study. The group composition is presented, followed by rating data and cut score outcomes, which are, in turn, followed by analyses of internal and external validity. Lastly, the survey and post-evaluation conference call data are presented.

A. Participants

Sixteen panelists participated in the research study. Two groups of eight panelists were balanced as closely as possible based on their primary specialty area of practice, practice setting, and time since initial certification. If possible, they were also matched as closely as possible with respect to gender, age, and geographical location. Several panelists had to be placed in either group 1 or group 2 due simply to availability. Ethnicity is reported but was unknown for most participants at the time of recruitment and group placement. Three individuals currently serving on the board of directors were split between the two groups (group 1: $n=1$, group 2: $n=2$). The composition of the groups is depicted in Table IV.
### TABLE IV

PANELIST COMPOSITION BY GROUP

<table>
<thead>
<tr>
<th>Panelist Characteristic</th>
<th>Category</th>
<th>Group 1</th>
<th>Group 2&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>White, not Hispanic or Latino</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Hispanic or Latino</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Black or African American</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Native American or American Indian</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Asian / Pacific Islander</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Did not specify</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Average Age (SD)</td>
<td></td>
<td>49.1 (8.28)</td>
<td>47.4 (5.34)</td>
</tr>
<tr>
<td>Geographical Location&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Northeast</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Southeast</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Midwest</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Southwest</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>West</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Primary Specialty Area</td>
<td>Comprehensive</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Practice</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cataract</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Glaucoma</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Neuro-Ophthalmology</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Retina</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Oculoplastics</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Uveitis</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Pediatrics</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cornea</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Oncology/Pathology</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Refractive</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Practice Setting</td>
<td>Academic</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Private practice</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Average Time Since Initial Certification (SD)</td>
<td></td>
<td>17.0 (8.87)</td>
<td>15.5 (4.85)</td>
</tr>
</tbody>
</table>
a Northeast includes Maine, Massachusetts, Rhode Island, Connecticut, New Hampshire, Vermont, New York, Pennsylvania, New Jersey, Delaware, and Maryland; Southeast includes West Virginia, Virginia, Kentucky, Tennessee, North Carolina, South Carolina, Georgia, Alabama, Mississippi, Arkansas, Louisiana, and Florida; Midwest includes Ohio, Indiana, Michigan, Illinois, Missouri, Wisconsin, Minnesota, Iowa, Kansas, Nebraska, South Dakota, and North Dakota; Southwest includes Texas, Oklahoma, New Mexico, and Arizona; West includes Colorado, Wyoming, Montana, Idaho, Washington, Oregon, Utah, Nevada, California, Alaska, and Hawaii.

b One group member participated in only one half of the study (item mapping method), and did not participate in the Angoff portion of the study due to changes in availability. In group 1, four individuals were considered early career (certified for 20 years or less). In group 2, seven individuals were considered early career.

B. Ratings and Cut Scores

1. Angoff

Group 1 participated in the Angoff method followed by item mapping, and group 2 participated in item mapping followed by the Angoff method. They rated the same 49 items for both activities. For all raters, average Angoff ratings were lower in round 2 than in round 1. The average ratings for all panelists are depicted in Table V.
TABLE V

ANGOFF RATINGS BY PANELIST

<table>
<thead>
<tr>
<th>Panelist</th>
<th>Round 1 Average Angoff Rating (SD)</th>
<th>Round 2 Average Angoff Rating (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>74.5 (11.8)</td>
<td>73.8 (13.9)</td>
</tr>
<tr>
<td>2</td>
<td>86.6 (12.5)</td>
<td>83.4 (12.9)</td>
</tr>
<tr>
<td>3</td>
<td>88.6 (5.3)</td>
<td>84.9 (13.0)</td>
</tr>
<tr>
<td>4</td>
<td>83.0 (7.0)</td>
<td>77.4 (13.6)</td>
</tr>
<tr>
<td>5</td>
<td>84.9 (9.9)</td>
<td>80.8 (14.5)</td>
</tr>
<tr>
<td>6</td>
<td>66.4 (15.7)</td>
<td>65.2 (15.2)</td>
</tr>
<tr>
<td>7</td>
<td>71.4 (12.9)</td>
<td>71.0 (15.6)</td>
</tr>
<tr>
<td>8</td>
<td>91.8 (6.2)</td>
<td>87.3 (10.4)</td>
</tr>
<tr>
<td>Group 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>80.2 (14.2)</td>
<td>79.2 (16.0)</td>
</tr>
<tr>
<td>2</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>3</td>
<td>82.6 (7.8)</td>
<td>78.4 (13.0)</td>
</tr>
<tr>
<td>4</td>
<td>87.6 (7.8)</td>
<td>84.5 (12.4)</td>
</tr>
<tr>
<td>5</td>
<td>71.9 (19.5)</td>
<td>64.2 (20.9)</td>
</tr>
<tr>
<td>6</td>
<td>79.7 (12.2)</td>
<td>77.1 (15.1)</td>
</tr>
<tr>
<td>7</td>
<td>83.7 (13.8)</td>
<td>82.1 (14.3)</td>
</tr>
<tr>
<td>8</td>
<td>76.6 (16.3)</td>
<td>74.1 (17.5)</td>
</tr>
</tbody>
</table>

*aPanelist 2 in Group 2 did not participate in the Angoff standard setting*

2. **Item mapping**

Group 2 participated in the item mapping standard setting first. They rated 26 of the 49 items before agreeing on a column in which they believed borderline examinees would answer exactly 50% of the items correctly. Group 1 rated 30 of the 49 items before agreeing on their cut score column. Because items were reviewed from easiest to hardest (and then the rating exercises stopped once the transitional column was identified), most of the items rated by both groups were easier items. The average p-value of the items rated by group 1 was 0.89, and the average p-value of the items rated by group 2 was 0.92. The average ratings for all panelists are depicted in Table VI.
### TABLE VI

<table>
<thead>
<tr>
<th>Panelist</th>
<th>Average rating (SD)</th>
<th>Item map column selected</th>
<th>Cut Score in logits (SE)</th>
<th>Raw Cut Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.85 (0.35)</td>
<td>18</td>
<td>0.91 (.34)</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>0.85 (0.38)</td>
<td>18</td>
<td>0.91 (.34)</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>0.88 (0.31)</td>
<td>18</td>
<td>0.91 (.34)</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>0.85 (0.35)</td>
<td>18</td>
<td>0.91 (.34)</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>1.00 (0.00)</td>
<td>19</td>
<td>1.02 (.34)</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>0.96 (0.18)</td>
<td>18</td>
<td>0.91 (.34)</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>0.96 (0.18)</td>
<td>18</td>
<td>0.91 (.34)</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>0.92 (0.25)</td>
<td>18</td>
<td>0.91 (.34)</td>
<td>33</td>
</tr>
<tr>
<td>Group 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.81 (0.40)</td>
<td>16</td>
<td>0.46 (.33)</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>0.83 (0.39)</td>
<td>15</td>
<td>0.35 (.33)</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>0.86 (0.36)</td>
<td>15</td>
<td>0.35 (.33)</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>0.85 (0.37)</td>
<td>15</td>
<td>0.35 (.33)</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>0.81 (0.40)</td>
<td>15</td>
<td>0.35 (.33)</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>0.77 (0.43)</td>
<td>15</td>
<td>0.35 (.33)</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>0.96 (0.20)</td>
<td>15</td>
<td>0.35 (.33)</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>0.88 (0.33)</td>
<td>15</td>
<td>0.35 (.33)</td>
<td>28</td>
</tr>
</tbody>
</table>

* Panelist 2 in group 2 joined the meeting late and only made 23 ratings; Panelist 3 in group 2 joined the meeting late and only made 14 ratings

3. **Cut scores**

The cut scores for the Angoff method were calculated by taking the mean of all panelists’ mean ratings for the 49 items. The cut scores for the item mapping method were calculated by identifying the column in which the majority of panelists agreed would result in the borderline examinee answering 50% correctly and transforming the logit value of the middle
item in that column to the raw cut score. The cut scores and corresponding pass rates for each condition are shown in Table VII.

**TABLE VII**

<table>
<thead>
<tr>
<th></th>
<th>Angoff</th>
<th>Item Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>Raw Cut Score (% correct)</td>
<td>39 (78%)</td>
<td>39 (78%)</td>
</tr>
<tr>
<td>Cut Score (logits)</td>
<td>1.67</td>
<td>1.67</td>
</tr>
<tr>
<td>Pass Rate</td>
<td>87.7%</td>
<td>87.7%</td>
</tr>
</tbody>
</table>

Both groups arrived at the same standard using the Angoff method but at different standards using the item mapping method. The item mapping standards were five raw questions apart (0.56 logits). This difference was compared using a z test for difference of item difficulty and was not statistically significant ($z = 1.18, p = 0.24$), but the item mapping standards were not combined into one cut score because it was hypothesized that the panelists might have different opinions about the acceptability of these two standards. Subsequent analyses and survey questions about the results compare three cut scores: the Angoff cut score, the item mapping group 1 cut score, and the item mapping group 2 cut score.

Table VIII lists the historical approved cut scores for this examination. Although these examination forms are not equated (new standards are set for each examination form), the
content outline was unchanged between 2014 and 2016, and there is no reason to believe the population of examinees would have changed significantly over this time period. Each of these examination forms had standards set using the Angoff method using 0-100 ratings.

TABLE VIII

HISTORICAL CUT SCORES FOR EXAMINATION

<table>
<thead>
<tr>
<th>Form Year</th>
<th>Cut Score (Pass Rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>38 (92.1%)</td>
</tr>
<tr>
<td>2015</td>
<td>39 (94.4%)</td>
</tr>
<tr>
<td>2014</td>
<td>41 (94.0%)</td>
</tr>
</tbody>
</table>

C. Intra-Panelist Consistency

For each panelist, round 2 Angoff ratings and item mapping ratings were correlated with empirical item difficulty values (p-values). Because the range is restricted with the item mapping rating data, correlations are lower than the Angoff ratings’ (0-100) correlation with p-values. To allow for a consistent comparison with the item mapping correlation data, Angoff ratings were transformed to 0/1 values (0-50 to 0 and 51-100 to 1). Correlations with both methods are depicted in Table IX.
### TABLE IX

#### INTRA-PANELIST CONSISTENCY

<table>
<thead>
<tr>
<th>Panelist</th>
<th>0-100 Correlation</th>
<th>0/1 Correlation</th>
<th>Item Mapping Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
<td>0.54</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
<td>0.54</td>
<td>0.41</td>
</tr>
<tr>
<td>4</td>
<td>0.97</td>
<td>0.66</td>
<td>0.54</td>
</tr>
<tr>
<td>5</td>
<td>0.90</td>
<td>0.50</td>
<td>n/a</td>
</tr>
<tr>
<td>6</td>
<td>0.76</td>
<td>0.55</td>
<td>0.12</td>
</tr>
<tr>
<td>7</td>
<td>0.84</td>
<td>0.62</td>
<td>0.33</td>
</tr>
<tr>
<td>8</td>
<td>0.95</td>
<td>0.54</td>
<td>0.40</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>0.92</td>
<td>0.58</td>
<td>0.44</td>
</tr>
<tr>
<td>Group 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.60</td>
<td>0.22</td>
</tr>
<tr>
<td>2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>n/a</td>
<td>n/a</td>
<td>0.06</td>
</tr>
<tr>
<td>3&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.97</td>
<td>0.61</td>
<td>0.40</td>
</tr>
<tr>
<td>4</td>
<td>0.96</td>
<td>0.61</td>
<td>0.23</td>
</tr>
<tr>
<td>5</td>
<td>0.71</td>
<td>0.62</td>
<td>0.33</td>
</tr>
<tr>
<td>6</td>
<td>0.86</td>
<td>0.61</td>
<td>0.41</td>
</tr>
<tr>
<td>7</td>
<td>0.97</td>
<td>0.60</td>
<td>-0.10</td>
</tr>
<tr>
<td>8</td>
<td>0.88</td>
<td>0.65</td>
<td>0.22</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>0.89</td>
<td>0.61</td>
<td>0.22</td>
</tr>
</tbody>
</table>

<sup>a</sup>Correlation could not be defined for panelist 5 because all ratings made were “yes” (“1”). There was no variance in ratings, which resulted in a denominator of zero in the correlation calculation.

<sup>b</sup>Panelist 5 was not used in calculation of overall average.

<sup>c</sup>,<sup>d</sup>Panelist 2 in group 2 joined the meeting late and only made 23 ratings; Panelist 3 in group 2 joined the meeting late and only made 14 ratings.

For both groups, the correlation between 0/1 Angoff ratings and \( p \)-values is greater than the correlation between item mapping ratings and \( p \)-values. Using a \( z \) test for difference of
correlations, the differences were not found to be statistically significant (group 1: \( z = 0.29 \) \((p = 0.38)\), group 2: \( z = 0.73 \) \((p = 0.23)\)).

D. **Inter-Panelist Consistency**

Inter-panelist consistency was assessed by comparing the SEJ for each condition (see Table X). Like the comparison of intra-panelist consistency, Angoff ratings were transformed to 0/1 for comparison. Inter-panelist consistency was highest for the group 2 item mapping standard, followed by the group 2 Angoff standard, the group 1 item mapping standard and, finally, the group 1 Angoff standard.

| TABLE X |
|------------------|------------------|
| INTER-PANELIST CONSISTENCY |                |
|                     | Angoff | Item Mapping |
| Group 1              | Group 2   | Group 1 | Group 2 |
| SEJ                  | 0.064    | 0.082   | 0.066   | 0.101   |

E. **Surveys**

The frequency of each response for each survey question is reported, as well as the median and interquartile ranges of the ratings on the score scale (with Strongly Agree = 4 and Strongly Disagree = 1).

1. **Initial training survey**

The first survey (Appendix B) was administered immediately following the initial training webinar. Eighty-seven point five percent (87.5%) of respondents strongly agreed that the
purpose and methods were clear and that they understood the definition of the borderline examinee. Eighty-seven point five percent (87.5%) of respondents strongly agreed or agreed that they were confident with applying the definition of the borderline examinee. No panelists entered any free-text comments. The frequencies of each response, median and interquartile range are reported in Table XI.

**TABLE XI**

<table>
<thead>
<tr>
<th></th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. Purpose and methods clear (n = 8)</td>
<td>7 ( (87.5%))</td>
<td>1 ( (12.5%))</td>
<td>1 ( (12.5%))</td>
<td>4.00</td>
<td>4.00-4.00</td>
<td></td>
</tr>
<tr>
<td>Q2. Understanding of borderline examinee (n = 8)</td>
<td>7 ( (87.5%))</td>
<td>1 ( (12.5%))</td>
<td>1 ( (12.5%))</td>
<td>4.00</td>
<td>4.00-4.00</td>
<td></td>
</tr>
<tr>
<td>Q3. Confident with applying definition of borderline (n = 8)</td>
<td>6 ( (75.0%))</td>
<td>1 ( (12.5%))</td>
<td>1 ( (12.5%))</td>
<td>4.00</td>
<td>4.00-4.00</td>
<td></td>
</tr>
</tbody>
</table>

2. **Angoff survey**

The survey that was administered after the panelists made Angoff ratings (Appendix C) is reported in Table XII. One hundred percent of respondents agreed or strongly agreed that they understood how to make Angoff ratings and that they were confident in their ability to apply the method. No free-text comments were provided that were relevant to the research questions or analysis.
TABLE XII

POST ANGOFF SURVEY

<table>
<thead>
<tr>
<th></th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. Understood how</td>
<td>6 (85.1%)</td>
<td>1 (14.3%)</td>
<td></td>
<td></td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td>to make Angoff</td>
<td>(n = 7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2. Confident in</td>
<td>4 (57.1%)</td>
<td>3 (42.9%)</td>
<td></td>
<td></td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td>ability to apply</td>
<td>(n = 7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(86.7%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. **Item mapping survey**

The survey that was administered after the item mapping method was conducted (Appendix E) is shown in Table XIII. Eighty-six point seven percent (86.7%) of respondents agreed or strongly agreed that they understood how to make their ratings and that they were confident in their ability to apply the method. One respondent reported that the item mapping standard setting was more intuitive than the Angoff standard setting. Two additional free-text comments were provided that were not relevant to the research questions or analysis.
The responses to questions 1 and 2 about the Angoff and item mapping methods in Tables XII and XIII were compared to determine if one method was rated as better understood (Question 1) or rated as having higher confidence in application (Question 2). Mann-Whitney U significance tests were used to compare the data for these two questions. The statement “I understood how to make my Angoff/item mapping ratings” did not have significantly different ratings between the methods ($U = 48.5$, $z = -0.25$, $p = .80$). The statement “I was confident in my ability to apply the Angoff/item mapping method” also did not have significantly different ratings between the methods ($U = 47$, $z = 0.35$, $p = .73$). However, of note, two individuals selected “Strongly disagree” in response to these two questions as related to the item mapping method. No respondents disagreed or strongly disagreed with these two questions as related to the Angoff method.

4. **Final results survey**

The results of the final survey administered to panelists (Appendix F) are reported in Table XIV and XV. Table XIV lists the results of the four Hofstee method questions.

### TABLE XIII

<table>
<thead>
<tr>
<th>Q1. Understood how to make item mapping ratings ($n = 15$)</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 (80.0%)</td>
<td>1 (6.67%)</td>
<td>2 (13.3%)</td>
<td></td>
<td>4.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q2. Confident in ability to apply ($n = 15$)</td>
<td>11 (73.3%)</td>
<td>2 (13.3%)</td>
<td>2 (13.3%)</td>
<td>4.00</td>
<td>3.00-</td>
<td>4.00</td>
</tr>
</tbody>
</table>
**TABLE XIV**

**HOFSTEE SURVEY RESULTS**

<table>
<thead>
<tr>
<th>Survey Questions</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. Avg. maximum acceptable passing score ($k_{max}$)</td>
<td>82.4 (9.8)$^a$</td>
</tr>
<tr>
<td>n = 11</td>
<td></td>
</tr>
<tr>
<td>Q2. Avg. minimum acceptable passing score ($k_{min}$)</td>
<td>70.6 (10.2)</td>
</tr>
<tr>
<td>n = 13</td>
<td></td>
</tr>
<tr>
<td>Q3. Avg. maximum acceptable failure rate ($f_{max}$)</td>
<td>14.6 (13.0)</td>
</tr>
<tr>
<td>n = 13</td>
<td></td>
</tr>
<tr>
<td>Q4. Avg. minimum acceptable failure rate ($f_{min}$)</td>
<td>3.2 (2.6)</td>
</tr>
<tr>
<td>n = 13</td>
<td></td>
</tr>
</tbody>
</table>

$^a$Two participants indicated that the maximum acceptable passing score was 100% in the survey. During the final results review, they both indicated that they misunderstood the question, so their responses were removed from the calculation.

The points (3.2, 82.4) and (14.6, 70.6) were plotted and intersected with the performance distribution at $x = 42.6$, resulting in a cut score of 84% correct (43 raw score points), as shown in Figure 3.
Figure 3. Hofstee method plot.

The Hofstee method cut score of 43 was compared with each cut score using a z test for difference of item difficulty. The Hofstee cut score was not significantly different from the Angoff cut score ($z = 1.12, p = 0.26$). However, it was significantly different from the group 1 item mapping cut score ($z = 2.54, p = 0.01$) and the group 2 item mapping cut score ($z = 3.58, p = 0.0003$).
The final survey questions asked whether panelists felt that each of the three cut scores were reasonable. The responses were coded as Strongly Agree = 4, Agree = 3, Disagree = 2 and Strongly Disagree = 1. A Kruskal-Wallis H test showed that there was no statistically significant difference in the ratings for the three survey questions relating to the reasonableness of the cut scores, \( \chi^2(2) = 3.26, p = 0.20 \), with a mean rank of 20.1 for the Angoff cut score, 23.7 for the group 1 item mapping cut score, and 16.3 for the group 2 item mapping cut score.

### TABLE XV

<table>
<thead>
<tr>
<th>Q1. Angoff cut score is reasonable ((n = 13))</th>
<th>Strongly agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly disagree</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>3.00</td>
<td>2.00-3.00</td>
<td></td>
</tr>
<tr>
<td>Q2. Group 1 item mapping cut score is reasonable ((n = 13))</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>3.00</td>
<td>2.00-4.00</td>
<td></td>
</tr>
<tr>
<td>(30.8%)</td>
<td>(23.1%)</td>
<td>(46.2%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3. Group 2 item mapping cut score is reasonable ((n = 13))</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>2.00</td>
<td>2.00-3.00</td>
<td></td>
</tr>
<tr>
<td>(30.8%)</td>
<td>(61.5%)</td>
<td>(7.69%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. **Post-rating evaluation conference call**

The purpose of the final conference call was to allow panelists to provide final input on the process. Twelve panelists participated in the call. A full transcript is provided (Appendix G), and nine key findings (and references to comments from panelists in support of those findings) are provided in Table XVI.
<table>
<thead>
<tr>
<th>Key finding</th>
<th>Corresponding lines in transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>Several participants had used the Angoff method previously and felt that</td>
<td>6-7, 9-11</td>
</tr>
<tr>
<td>they were more comfortable using it because of their previous experience.</td>
<td></td>
</tr>
<tr>
<td>Some panelists thought that their subspecialty area of expertise might</td>
<td>32-33, 35-37</td>
</tr>
<tr>
<td>affect their item ratings. Two panelists felt that they gave higher ratings</td>
<td></td>
</tr>
<tr>
<td>to items in their subspecialty area because they judged them as easier.</td>
<td>58-62, 64-69, 509, 513</td>
</tr>
<tr>
<td>Some panelists found that making a yes/no judgment (in the item mapping</td>
<td></td>
</tr>
<tr>
<td>procedure) was more challenging than selecting a value between 0 and 100 (in</td>
<td></td>
</tr>
<tr>
<td>the Angoff procedure). Some panelists had trouble understanding the four</td>
<td></td>
</tr>
<tr>
<td>questions relating to the Hofstee method. Two individuals mentioned that</td>
<td>115-119, 126, 128</td>
</tr>
<tr>
<td>they misinterpreted the term “maximum acceptable passing score” as “maximum</td>
<td></td>
</tr>
<tr>
<td>acceptable pass rate.” Others indicated that they had to think a lot about</td>
<td></td>
</tr>
<tr>
<td>these questions before they understood what they meant.</td>
<td>257-267, 271-272, 339-340, 387-389,</td>
</tr>
<tr>
<td></td>
<td>425, 433-437, 466-474</td>
</tr>
<tr>
<td>It was challenging for the panel to agree on a pass rate that seemed</td>
<td></td>
</tr>
<tr>
<td>acceptable because the population is highly subspecialized. This may</td>
<td></td>
</tr>
<tr>
<td>have contributed to some of the diversity in the ratings of each cut score’s</td>
<td></td>
</tr>
<tr>
<td>acceptability. The panelists felt that the passing standard set by the</td>
<td>304-306, 312-314, 314, 320-322,</td>
</tr>
<tr>
<td>Angoff method seemed appropriate in terms of the number of items needed to</td>
<td>390-393, 397-398, 415-417, 419,</td>
</tr>
<tr>
<td>answer correctly but felt that the corresponding number of people who would</td>
<td>421, 488-491, 528, 530, 532, 536-</td>
</tr>
<tr>
<td>fail if that standard was implemented was too high. The group was divided</td>
<td>539-544, 546, 551</td>
</tr>
<tr>
<td>on whether the group 1 item mapping cut score was acceptable.</td>
<td></td>
</tr>
<tr>
<td>The group agreed that the group 2 item mapping cut score was not</td>
<td></td>
</tr>
<tr>
<td>acceptable. The group was more vocal about preferring the process of the</td>
<td></td>
</tr>
<tr>
<td>Angoff method although some still expressed dissatisfaction with the</td>
<td></td>
</tr>
<tr>
<td>resulting pass rate.</td>
<td></td>
</tr>
</tbody>
</table>
III. DISCUSSION

This research evaluates the validity arguments of the Angoff and item mapping standard setting methods with application to a high-stakes recertification examination. This chapter reviews each method with respect to procedural validity, internal validity, and external validity. In addition, limitations of the interpretations and suggestions for future research are presented.

A. Procedural Validity

Evidence for procedural validity includes documentation that all steps in the standard setting process were carried out appropriately: a sufficient number of panelists engaged in the standard setting, the panelists are representative of the population of certificate holders, and the standard setting purpose and training was understood by the panelists. Evaluation of procedural validity is often overlooked or assumed, but is essential to establish. When lacking, it has the potential to invalidate a standard (Kane, 1994). Much of the procedural validity for this standard setting was established as part of the study design. Chapter 3 outlines the research on the ideal number of standard setting panelists, and efforts were made in the recruitment process to secure at least six individuals per panel. The study was also designed to maximize the representativeness of the panelists of the population of certificate holders and assess their understanding of the training and rating tasks through online surveys.

1. Panelist representation

An important element of procedural validity is appropriate panelist representation. A random sample of certificate holders to participate in the standard setting would be ideal, but is not often feasible in practice. For this study, recruitment involved asking item-writers of the certification board to volunteer for the standard setting. It is possible that this group of volunteers had different characteristics than the overall group of certificate holders. For example, the
volunteers could be more likely to be high-achieving or to be in practice settings with more free time to allow such participation. This impacts the argument for procedural validity although it is not uncommon in standard setting studies.

The group was well-represented in terms of diversity in practice setting, age, gender, and primary specialty area of practice. The panelists were split into groups based in part on availability, which resulted in some imbalances in the groups. First, group 1 had three comprehensive practitioners represented, whereas no comprehensive practitioners were members of group 2. It is possible that comprehensive practitioners have a different approach or perspective than subspecialists, and this could have contributed to their ratings. Although at least one study has found that subspecialists rate similarly to non-subspecialists (Norcini, Shea, & Kanya, 1998), at least some anecdotal evidence gathered in the post-meeting evaluation conference call suggested that subspecialists rate items from their subspecialty area of expertise differently. More research is needed to determine if/what effect subspecialization has on ratings.

Secondly, group 1 had more academic practitioners (75% in group 1, 62.5% in group 2). While the exact distribution of certificate holders in the United States in academia and in private practice is unknown, including more private practitioners as panelists would likely provide a more representative sample. Additionally, the groups were not balanced in terms of time since initial certification. Although the average ages of the panel members were similar (49.1 for group 1, 47.4 for group 2), only four individuals had been in practice for 20 years or less from group 1 as compared with seven individuals in group 2. It is possible that individuals with different amounts of experience since obtaining initial certification would have different views about the strengths and weaknesses of a borderline examinee. It is also possible that the experience of panelists with actual examinees that they would consider borderline may have affected their
ratings. It is unknown whether this had an impact, and if so, whether certain panelists (e.g., private practitioners, those in practice a longer time, etc.) may have been more likely to have been affected by such experience. Lastly, one goal of the initial design was inclusion of at least one panelist per subspecialty area in each group. No cataract subspecialists, refractive subspecialists, uveitis subspecialists, or oncology/pathology subspecialists volunteered to participate. With additional insight from individuals from these subspecialty areas, it is possible the group definition of minimal competence would have been different or that some items would have been rated differently.

2. **Procedural understanding**

Procedural validity was also evaluated by analyzing how clearly the standard setting processes and procedures were explained and understood through anonymous survey questions. Eighty-seven point five percent (87.5%) of respondents (7 out of 8) indicated that they strongly agreed that the training of the standard setting purpose and methods were clear (Appendix B, question 1), and that they understood the definition of the borderline examinee (Appendix B, question 2). The same number agreed or strongly agreed that they were confident in their ability to apply the definition of the borderline examinee (Appendix B, question 3). One limitation to these findings is the small number of respondents. Even though all 16 panelists were asked to complete the survey, only eight did so. It is possible that the eight panelists who did not complete the survey would have responded differently.

Additionally, procedural validity was assessed through panelist confidence in the application of the methods (making ratings). There was not a significant difference in the survey results for the Angoff and item mapping methods with respect to panelists’ understanding of how to make their ratings and their confidence in their ability to apply the method (Appendices C and
During the post-meeting evaluation conference call, panelists were asked to comment further on these questions. Regarding the Angoff method, panelists expressed that they understood how to make Angoff ratings and were confident that they understood the task, but they were still surprised at times when the actual performance data \((p\text{-values})\) were presented. In other words, they felt confident in coming up with a rating, but when that rating was largely different from the empirical \(p\)-value, it was challenging to reconcile. The relationship between ratings and \(p\)-values is explored further in the section on internal validity.

With respect to the item mapping method, panelists understood how to make their ratings, but they expressed less confidence in making “yes” or “no” judgments. One panelist referred to this complexity as a “coin toss” (Appendix G, line 58), and another indicated it was “hard to create that line” (Appendix G, line 68), which means that, for some items, it was hard to decide if the borderline examinee would answer correctly or not. One panelist mentioned that the item map biased his/her judgments. Because the items were lined up from easiest to hardest, the panelist had trouble making a judgment on the likelihood of a borderline examinee answering correctly based on judgment alone because the location of the item on the item map was “hard to ignore” (Appendix G, line 56).

Overall, the panelists expressed understanding in both methods and confidence in their ability to apply the methods. The aforementioned issue related to incongruency between perceived and actual item difficulty requires additional attention and research.

**B. Internal Validity**

Internal validity was assessed through intra- and inter-panelist consistency, both of which contribute to the reliability of the cut score. Reliability alone does not make for validity, but, when lacking, it detracts from the validity argument of the standard.
1. **Intra-panelist consistency**

Intra-panelist consistency was measured through correlations between ratings and empirical \( p \)-values. When the relationship between ratings and empirical \( p \)-values is strong, we might conclude that the panelists are able to consistently apply the definition of the borderline examinee. The correlations between 0-100 Angoff ratings and \( p \)-values were strong (.92 for group 1, and .89 for group 2). When transformed to 0/1 ratings, the correlations were .58 for group 1 and .61 for group 2, respectively. For the item mapping method, the ratings’ correlations with \( p \)-values were .44 and .22 for groups 1 and 2, respectively. Although these differences were not statistically significant for either group, the correlations were stronger for the Angoff method.

Consistent application of a shared definition of minimal competence would theoretically result in ratings that correlate with empirical \( p \)-values if we assume that MCE and overall group performance is correlated (Van der Linden, 1982). In fact, if a group of panelists had a shared conceptualization of what a borderline examinee would and would not answer correctly (and the ability to apply that knowledge to an item), we’d expect to see both high intra- and inter-panelist consistency. However, such agreement could be the result of panelists over-relying on external data. Because \( p \)-values were provided to panelists for all items before making their round 2 Angoff ratings, it could be the case that panelists simply adjusted their ratings to better correspond with the \( p \)-value for the item.

When evaluating the item mapping method with respect to intra-panelist consistency, two things are worth noting. First, the item mapping method required panelists to rate items until they identified a transitional column, which meant that the majority of items being rated were easier items. Plake, Impara, and Irwin (2000) found that panelists are less consistent when rating easier
items. Secondly, $p$-values were not provided to panelists unless they disagreed about their ratings for an item. The item map showed the relative difficulty of the items, but it is possible the panelists’ ratings would have higher intra-panelist consistency if they had seen the $p$-values for all items or if all items (easy and hard) were rated. For future research, it might be advisable to encourage panelists to make ratings several columns beyond where they initially believe the transitional column to be. The review of harder items may impact their perception of the difficulty of all items, and ultimately impact the resulting cut score.

For both methods, one limitation to the finding of no significance was that the number of panelists for each group was small ($N = 7$ and $N = 8$ for groups 1 and 2, respectively).

2. **Inter-panelist consistency**

Inter-panelist consistency is a measure of the consistency of ratings across panelists. It is an important aspect of the validity argument, as it indicates whether we’d expect similar results if the study were repeated with different panelists. The SEJs for each method’s ratings suggest that panelists were relatively similar to one another in terms of their ratings.

One limitation to this interpretation is that the Angoff data was transformed to 0/1 to allow for a consistent comparison between the Angoff and item mapping ratings. This dichotomization of the judgments made on a continuous scale takes away the precision of the estimates made by the panelists. Systematic variability is lost, and lower variability contributes to low correlations. A second limitation is that the item mapping ratings were made on a subset of the examination items that were different in difficulty than the examination form overall. It is likely that more variance in ratings and a lower SEJ would have resulted if the item mapping method had involved rating all 49 items. An analysis that was not completed was an evaluation of each panelist’s Angoff standard as compared to the others’ standards. In some cases, when a
panelist’s average rating is a specified number of standard deviations away from the mean overall rating, a cut score might be calculated both with and without the ratings of said panelist. Thus, a limitation to this study is that the potential influence of outlier panelists was not investigated.

Further support for the reliability of the Angoff method is in the comparison of the cut score outcomes across panels. Both groups independently arrived at the same cut score using the Angoff method. For the item mapping method, although the cut scores were not statistically significantly different, the groups arrived at cut scores that were different by five raw points out of 50.

C. External Validity

1. Hofstee method comparison

The Hofstee standard setting method was conducted and the resulting cut score was compared to the Angoff and item mapping standards. When two methods for standard setting result in a similar standard, the external validity argument of that standard is strengthened.

Using survey data from panelists (Appendix F, questions 1-4), the Hofstee method standard setting resulted in a raw cut score of 43, which was not significantly different from the Angoff or group 1 item mapping cut scores (39 and 33, respectively) but is significantly different than the group 2 item mapping cut score (28). Key finding #4 from the post-evaluation conference call discussion outlines the challenge panelists had in answering the four survey questions related to the Hofstee method. Two panelists noted during the call that they misunderstood at least one of the four questions completely, confusing the concepts of pass rate and passing standard. Although these two data points were removed from the final calculation
for the Hofstee method, it is unclear how reliable the remainder of the data is. It is also unlikely that the group would find the Hofstee method’s result to be reasonable (a cut score of 43 results in a pass rate of 54.5%) since many agreed that the Angoff cut score of 39 was close to being too high. For purposes of validating the cut scores by comparison to the Hofstee method cut score, much discretion should be used in interpretation given the potential lack of understanding by the panelists.

2. **Standard reasonableness**

Despite unclear results about the reasonableness of the Angoff and item mapping cut scores through comparison to the Hofstee cut score, the reasonableness of the standards were assessed through survey questions and additional discussion among panelists. Panelists with a good understanding of the knowledge, skills, and abilities of the pool of certificate holders can provide insight into the reasonableness of a standard. Despite all attempts at implementing a sound standard, standard implementation is ultimately a policy decision. If implementation of a standard results in a pass rate that is considered to be too low or too high to the decision-making body, it is likely to be rejected.

The task of assessing the reasonableness of a standard was operationalized by asking panelists if the cut scores for each method and resulting pass rates were reasonable. The results (shown in Table XV) indicate disagreement among the panelists. While some panelists thought the Angoff standard was the most reasonable, others felt the group 1 item mapping standard was most reasonable. While the results of these survey questions were not statistically significantly different, the median response for the Angoff and group 1 item mapping cut scores (median = 3) was greater than the median response for the group 2 items mapping cut score (median = 2).
Additionally, the group 2 item mapping cut score is the only cut score for which any panelists strongly agreed that the standard was reasonable ($n = 4$).

Historically, the pass rate for this examination module has been between 92 and 94.5% (between 2014 and 2016). The cut score that was actually implemented for this examination form (using the Angoff method) was 38 (pass rate of 92.1%). This was consistent with the discussion among the panelists, who expressed some discomfort with the “low” pass rate corresponding to the Angoff method cut score (at a cut score of 39, the pass rate would be 87.7%).

In the post-meeting evaluation conference call, the panelists discussed the reasonableness of each standard. They agreed that the group 2 item mapping standard (29) was too low.

D. **Summary**

In the context of standard validation, what matters more - the process, or the outcome? Both contribute to the validity argument of the standard. The Angoff method process was preferred by most panelists, but the outcome of the group 1 item mapping standard was preferred over the Angoff method by several panelists. On the other hand, the group unanimously agreed that the group 2 item mapping standard was too low. In this case, neither standard setting method clearly met standards for procedural, internal, and external validity better than the other.

However, this research contributes significant new information to the field related to these standard setting methods, and the relative strengths and weaknesses of each method should be considered in application to different examinations.

The item mapping method showed promise in several key areas. Panelists indicated that they were confident applying the item mapping method. Also, they did not report perceptions of item disordinality (which has been cited as an issue with the traditional bookmark method (Hein
& Skaggs, 2009)). Both groups were able to identify a transitional column where they felt a borderline examinee would have a 50% chance of answering the items correctly. One panelist reported that the item mapping method was “more intuitive” than the Angoff method.

However, the item mapping method was less effective in other areas. One area of issue was that the two groups arrived at different standards using the item mapping method. This indicates a lack of internal consistency. Another issue was that some panelists were more confident making 0-100 ratings as opposed to yes/no ratings. Wang (2003) found that the item mapping method resulted in higher inter-panelist consistency than the Angoff method, but, in this study, no statistically significant difference was found.

Another possible shortcoming of the item mapping method is revealed by the implication that the ratings may have been influenced by other panelists. With the item mapping method, ratings are announced aloud. If there is disagreement, discussion about the item occurs and then panelists re-rate. It is possible that this group dynamic discouraged disagreement in the minds of some panelists. No panelists specifically mentioned that this occurred, but at least one mentioned that they preferred the Angoff method due to the “isolation.”

There are some limitations to the generalizability of the item mapping results given the specific implementation practices that were used in this study. The study did not include a final step in the traditional item mapping method procedure. To allow for consistent comparison between the Angoff and item mapping methods, no impact data was provided with either method. This allowed for evaluation of the two methods only in terms of their validity arguments using criterion-referenced data. The introduction of norm-referenced information (impact analysis or provision of conditional pass rates) was hypothesized to impact the ultimate cut scores in a way that could potentially obscure the information gathered by ratings alone. Such
information has the potential to be “hugely influential” on ratings and, thus, on cut scores (Skorupski, 2012, p. 141). The Hofstee method was introduced as the sole standard that was set which incorporated information about norms (acceptable pass and fail rates). The traditional item mapping method utilizes a probability ruler, whereby the pass rate associated with a cut at a particular column is made known to panelists during the standard setting, and they are given an option to revise their ratings and decision about the transitional column. Had this step been included, it is possible that the groups would have agreed upon standards that they felt were more reasonable.

The Angoff method had several positive outcomes. First and foremost, six panelists vocalized their ultimate preference for the Angoff method over the item mapping method at the conclusion of the study (after having seen the resulting cut scores and corresponding pass rates). The panelists rated the Angoff highly with respect to their understanding of the method and confidence in their ability to apply it. Additionally, both groups arrived at the same cut score using the Angoff method, which is evidence of strong reliability.

The Angoff method, however, had some weaknesses. First, it is important to note that some panelists and the facilitator had experience in setting standards using the Angoff method. None of the panelists or the facilitator had experience with the item mapping method. Some panelists may have been more comfortable using the Angoff method due to their history of engaging in this type of rating task. Panelists with experience using the Angoff method may have been more likely to prefer the 0-100 ratings to the yes/no ratings simply because they were accustomed to that model. Additionally, the Angoff method resulted in a cut score that some panelists felt was too high. This could be based on their knowledge of the examinee population, on their knowledge of historical pass rates for this examination, or both.
There were some limitations that affected this study as a whole that should be considered. One issue was the disagreement among panelists about the true need for a common standard on this examination. Some panelists felt that the core knowledge needed by certificate holders was not represented by the knowledge being tested on this examination. The panelists mentioned different standards for different types of specialists or different, customized examinations for use in recertification. Without a clear test purpose and agreement on the knowledge, skills, and abilities needed for recertification, attempts at identifying an appropriate standard are extremely challenging. A second issue was that the study was conducted remotely. Most standard setting studies include in-person training and discussion. It is possible that the remote environment made panelists less likely to be engaged in the discussions or less likely to ask questions. Additionally, due to the remote environment, the study was conducted over a five-week period in part because was hypothesized that extending this timeframe any longer would result in attrition of panelists. However, it is possible that a longer timeframe that allowed for provision of some feedback between the first and second rounds of Angoff ratings would have been beneficial and resulted in a different cut score. Thirdly, the environmental differences between the Angoff and item mapping methods may have contributed to the results. The item mapping method took place on a group conference call. Two panelists joined the group 2 meeting late, which may have impacted their ratings or the ratings of others. Additionally, the item mapping conference calls were scheduled for 90 minutes. Both groups used the full 90 minutes. Although both agreed on an item mapping cut column in that allotted time, it is possible that more time would have led to rating of additional items and deciding on a different cut point. One panelist mentioned that he/she thought that time might have impacted the item mapping cut score. On the other hand, the Angoff method was completed by individuals independently during an approximately two-week
window. On one hand, this allowed individuals to engage in the activity at their convenience. On the other hand, because of the uncontrolled environment, it is unknown whether additional factors affected their ratings. For example, it is possible that individuals who completed their ratings earlier in the window had better recollection of the initial training. Lastly, the ability to evaluate the reasonableness of the pass rate resulting from the cut score could have been affected by the panelists’ knowledge of the historical pass rates for the examination. The fact that the pass rate for the entire three-part examination had never previously been below 95% could have contributed to the perception of the Angoff standard’s pass rate (87.7%) as being “too low.”

E. Significance and Practical Implications

This study provides the research community with important considerations for future standard setting research. First, the finding that the panelists were more comfortable making a judgment on a 0-100 scale as opposed to yes/no ratings conflicts with previous findings (Impara & Plake, 1997). Additional studies comparing methods using 0-100 and yes/no ratings with respect to procedural, internal, and external validity would be beneficial. Second, the item mapping method showed promise, and more research should be dedicated to its evaluation. Based upon the outcomes of this study, the final step in the item mapping method (impact information) should be included for a complete evaluation of the method as it was intended to be implemented. Also, the impact of using a different RP than .50 for the item mapping method would provide valuable information to the field. It is possible that if panelists conceptualized an individual that had mastered 60% or 70% of the content, the method would have resulted in a cut score that was found to be more acceptable. Additionally, this research did not evaluate the standard setting methods with respect to time, cost, or efficiency. Studies on the amount of time needed for sufficient training, time needed for different methods, and efficacy of remote standard
setting would provide beneficial data to operational programs looking to balance cost-effectiveness and validity. Lastly, this study was one of only a handful that engaged panelists in an extended post-meeting discussion. The results of the post-meeting evaluation conference call provided insights that were not articulated through the online surveys. The current norm of establishing procedural validity by administration of post-meeting surveys alone should be re-evaluated, and additional studies seeking qualitative feedback from panelists should be conducted. It is essential that we continue to investigate the cognitive processes of panelists in future standard setting research.

For purposes of conducting future standard settings, I would recommend testing organizations consider utilizing a standard setting method that combines elements of both studied methods. In this study, panelists preferred making 0-100 ratings. I also believe they benefited from the group discussion that occurred between rounds during the item mapping method, but ideally then would have made their final ratings without seeing the ratings from the other panelists. Also, these methods were compared without between-round feedback for the Angoff method (e.g., a rank-ordering of standards as set by each panelists), which may have provided additional insight to the panelists. I recommend that additional research be done to evaluate the impact of different types of feedback and at different times throughout the standard setting exercises.

Standard setting “… is one of the most complex, consequential, and vexing issues facing specialists and policy makers today” (Cizek, 2012a, introduction page). It is imperative that the measurement community continue to evaluate standard setting methods and practices. Despite the litany of research on standard setting, many questions remain unanswered.
REFERENCES


standard setting using two variations of the bookmark method. *Applied Measurement in
Education, 22*(3), 207-228. doi: 10.1080/08957340902983978

Anderson and J. S. Helmick (Eds.), *On educational testing* (pp. 109-127). San Francisco,


Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the
assumptions in the Angoff standard setting method. *Journal of Educational

School Competency Test*. Paper presented at the spring meeting of the North Carolina
Association for Research in Education, Chapel Hill, NC.

competency tests: Theory and application. *Educational Evaluation and Policy Analysis,


value of observed behaviors on military job performance tests. *Performance assessment


Linacre, J. M. (2016). Winsteps (Version 4.0.0 ) [Computer software]. Chicago, IL: Winsteps.com


doi:10.1207/s15324818ame1803_5


Appendix A – Borderline Definition

Examples of borderline ophthalmologist characteristics:

- Cataract patient – exam assess for phacodonesis, didn’t notice lens had movement
- Able to recognize a retinal detachment and send to retina specialist
- Recognizes diabetic retinopathy but may not be able to treat, recognizes urgency
- Recognizes basic horizontal strabismus but does not have a treatment plan...would refer.
- Recognizes possible retinoblastoma and importance of quickly referring
- Able to diagnose a corneal ulcer but unable to identify/treat uncommon causes including fungal and acanthamoeba keratitis
- Patient followed as a glaucoma suspect, but ophthalmologist doesn't recognize progressive visual field loss until more advanced
- A candidate who knows standard treatments but is not aware of new or experimental treatments that are available and would be appropriate for an individual patient
- Recognizes a BRVO or CRVO but doesn't counsel patient about neovascular glaucoma risk
- Ability to properly evaluate glaucoma progression with OCT and VF
- Recognizes basic lacrimal duct obstruction and understands the timeline of surgical treatment
- Recognize early signs of post-operative endophthalmitis
- Variable ptosis associated with diplopia could indicate neuromuscular junction disorder and could be associated with systemic respiratory failure
- Patient with a CRAO, no plaque, not does not order ESR/CRP/platelets to rule out other causes like giant cell arteritis
- Understanding the risk of Plaques toxicity and the increasing risk with duration of treatment and cumulative dose
- Able to diagnose giant cell arteritis prior to second eye involvement but not necessarily able to treat
- Patient with ocular trauma, subconjunctival hemorrhage, but does not dilate to rule out other trauma like RD
- Floppy eyelids and lash ptosis may be associated with obstructive sleep apnea
- Can diagnose keratoconus, but does not know all the surgical and non-surgical treatments
- Patient with nonresolving uveitis, does inquire about past history of ocular lymphoma
- Recognizing amblyopia in the age group up to age 10 and knowing how to treat or recognizing importance of referral
Appendix B – Survey Administered After Initial Training

Directions: Please indicate your agreement with the following statements pertaining to the standard setting initial training. Your responses and any comments will remain anonymous.

1. Overall, the training in the standard setting purpose and methods was clear.
   Strongly Disagree  Disagree  Agree  Strongly Agree

2. I have a good understanding of the definition of the borderline examinee.
   Strongly Disagree  Disagree  Agree  Strongly Agree

3. I am comfortable with my ability to apply the definition of the borderline examinee to examination items.
   Strongly Disagree  Disagree  Agree  Strongly Agree

4. Additional comments (optional):
Appendix C – Survey Administered After Both Rounds of Angoff Ratings

Directions: Please indicate your agreement with the following statement pertaining to the Angoff standard setting. Your response and any comments will remain anonymous.

1. I understood how to make my Angoff ratings.
   Strongly Disagree   Disagree   Agree   Strongly Agree

2. Overall, I am confident that I was able to apply the Angoff standard setting method appropriately.
   Strongly Disagree   Disagree   Agree   Strongly Agree

3. Additional comments (optional):
Appendix D – Item Map

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
|   |   |   |   |   |   |   |   |   |   | 5325 | 4467 | 4538 | 4392 | 4234 | 4136 | 4240 | 2911 | 1729 | 1567 | 6513 | 9014 | 15129 | 4541 | 4113 | 6475 | 6475 | 6475 |
| 1 |   |   |   |   |   |   |   |   |   | 6670 | 4435 | 4380 | 10801 | 6540 | 6660 | 10732 | 7804 | 5871 | 3166 | 4136 | 3720 | 4671 | 5417 | 3877 | 3971 | 75 | 5637 | 6595 | 5364 | 6475 |

**Easy**

**Hard**

**Color Key:**

- Cataract and Anterior Segment
- Cornea and External Disease
- Glaucoma
- Neuro-Ophthalmology and Orbit
- Oculoplastics and Orbit
- Ophthalmic Pathology and Oncology
- Pediatric Ophthalmology and Strabismus
- Refractive Management and Optics
- Retina and Vitreous
- Uveitis
Appendix E – Survey Administered After Item Mapping Ratings

Directions: Please indicate your agreement with the following statement pertaining to the item mapping standard setting. Your response and any comments will remain anonymous.

1. I understood how to make my item mapping ratings.
   Strongly Disagree  Disagree  Agree  Strongly Agree

2. Overall, I am confident that I was able to apply the item mapping standard setting method appropriately.
   Strongly Disagree  Disagree  Agree  Strongly Agree

3. Additional comments (optional):
Appendix F – Final Survey

Directions: The four questions below ask for your judgment about the range of acceptable pass rates and scores on this examination. Your responses and comments will remain anonymous.

1. What do you think is the maximum acceptable passing score (percent correct needed to pass) on this module? E.g., you may think requiring an examinee to answer 95% of questions correctly in order to pass is too high of a standard. In this case, you might designate the maximum acceptable passing score as 90%.

2. What do you think is the minimum acceptable passing score on this module? E.g., you may think requiring an examinee to answer 60% of questions correctly in order to pass is too low of a standard. In this case, you might designate the minimum acceptable passing score as 65%.

3. Given your knowledge of the qualifications and performance of the examinees that take this module, what do you think is the maximum failure rate that would be acceptable for this module? E.g., is it acceptable if 30% of individuals fail? 40%?

4. Given your knowledge of the qualifications and performance of the examinees that take this module, what do you think is the minimum failure rate that would be acceptable for this module? E.g., is it acceptable if 5% of individuals fail? 0%?

Directions: Listed below are the cut scores and corresponding pass rates for each standard setting method (Angoff and item mapping). Please indicate your agreement with the final statements pertaining to your judgment of the acceptability of each of these cut scores.

After submitting responses to questions 1-4 above, cut scores and corresponding pass rates for Angoff and item mapping were listed. Respondents were unable to go back and edit their responses to questions 1-4 once the cut scores and pass rate information was provided.

5. The final recommended cut score using the Angoff method (39 out of 50) is reasonable (fairly represents the minimum level of acceptable performance).

Strongly Disagree  Disagree  Agree  Strongly Agree

6. Additional comments (optional):
Appendix F – (continued)

7. The final recommended cut score from Group 1 using the item mapping method (33 out of 50) is reasonable (fairly represents the minimum level of acceptable performance).

   Strongly Disagree  Disagree  Agree  Strongly Agree

8. The final recommended cut score from Group 2 using the item mapping method (28 out of 50) is reasonable (fairly represents the minimum level of acceptable performance).

   Strongly Disagree  Disagree  Agree  Strongly Agree

9. Additional comments (optional):
Appendix G – Transcript of Final Results Conference Call

Not listed in the transcription are comments that were not directly related to the interview (e.g., logistics for logging on/viewing the screen/muting), identifying information (individuals often introduced themselves before making a comment), repeated words or non-lexical conversation sounds (e.g., um) and comments that were off topic or unrelated to the questions asked in the focus group (e.g., questions or comments about the initial certification exam(s) or item development.)

Panelists are identified by group and rater ID number. For example, rater 2 in group 1 is listed as G1-2. The facilitator and panelists were typically referred to by one another by first name, but they are coded by panelist ID and as “facilitator” in the transcript.
The attendees were asked for feedback relating to these two survey statements:

- I understood how to make my Angoff ratings.
- Overall I am confident that I was able to apply the Angoff standard setting method appropriately.

G2-6: Had done Angoff before, so I was familiar and comfortable with it, so, I don’t know how many other people are in that same boat, but-

G1-4: I felt the same way [comfortable with Angoff because I’ve done it before]. I understood the task and that we were looking at the borderline examinee - not just everybody, because I think that’s where people get confused sometimes.

G1-7: I never made Angoff ratings before and I felt it was pretty straightforward on how to make them and I had to refresh myself of the definition of the borderline examinee as I was going throughout my ratings, and that was helpful.

G2-1: So I have done the Angoff before so I felt more comfortable with it, but still there were some items that caught me by surprise and I put those in the comments. And the other thing was I was more used to doing things in isolation rather than in a group, which is what the other technique was, so I was more comfortable with being in isolation.

G1-2: I’ve done the Angoff before so I think I understand the idea of how to do it but I still sometimes feel like I’m just taking a stab in the dark when I do that, because I too was surprised sometimes when we got the actual statistics of a previously asked question as to how many people had gotten it correct or incorrect.

G1-1: I agree with G1-2. I feel like I have a pretty good handle on how to Angoff but I still, there are still some questions where I think I’ve gotten a pretty good handle on how many individuals will get it right or wrong and then the statistics are shown and I’m just way off – and then some I feel pretty good, I have a pretty good feeling for that. And I feel like the more I do the Angoffing the better I am but there are some questions that catch me by surprise, I always wonder whether it is a bit specialty-specific – I feel like I expect more knowledge about my own specialty than about other specialties, and so that may affect my choice of score sometimes.

G1-8: I agree with that [my specialty area affects my ratings]. I feel like especially when I’m grading my own specialty I tend to feel like ‘that’s common knowledge, everybody knows that’, so my Angoffs are a little bit skewed, kind of.

G1-6: I agree with both of those comments. I think it’s not only specialty-based but also a test-taking base, so, if you’re- I don’t know if this can be consciously or subconsciously done but if you’re a good test taker you might score more highly and if you’re not a great test-taker less [one second inaudible]… also your impression of how people should be doing on the test based off your own personal experience.
The attendees were asked for feedback relating to these two survey statements:

- I understood how to make my item mapping ratings.
- Overall I am confident that I was able to apply the item mapping standard setting method appropriately.

G2-1: I think the visual impact of the item mapping kept creeping into my bias, so, easiest were on the left, harder on the right, so I had some idea of where the questions stood whereas in the Angoff, since we’re not given past performance data, I felt like ok, I’m seeing this item in isolation with no, other than just intuiting whether I thought it was hard or difficult, I’m sorry easy or difficult, that it was the first round was very truly ‘what do I think’ without the bias of seeing it on that visual scale where it’s in the middle then it’s kind of ‘medium-hard’ and that sort of thing, so I did have sense that I was biased in that way so I’d set an expectation, and I thought if you realize that, you can overcome the bias or do something, that was hard to ignore.

G1-4: I had the same, I feel like 50/50 was such a coin toss and knowing that there aren’t many borderline examinees, I think it was hard to say ‘ok, well, 50% of them are going to get this wrong,’ when, like where to find that exact line of where that’s going to happen. You obviously don’t want to make it way way up high, but I felt like it was a little tough to blend that with it being 50/50.

G1-1: I was struck by the diversity of responses when we were trying to find that 50/50 mark as well, Facilitator, it is hard to identify that line, and when we are going through the questions and deciding whether it is 50% would get it right or 50% would get it wrong, most of the folks on our panel, we had difficulty coming to any resolution, so, and I think even looking at the data that we had it was hard to create that line, that 50/50 line, it did seem like a rather broad decision to make and I think that’s maybe why we struggled to make, to reach agreement.

G1-2: It would have been interesting to map some that hadn’t previously been given because it’s hard to not be influenced by the previous track record of the question, and yet we had to do that with some Angoffing in the past.

Facilitator: Can you remind me were you in group 1 or group 2? Did you do item mapping and then Angoff, or-

G1-2: I was in group 1 so I did Angoff first and then item mapping.

Facilitator: Do you think that that had an impact on you? I think I’m hearing you say yes, that you saw those items already, you Angoffed them, and then went onto item mapping and that may have influenced your ratings in item mapping?

G1-2: I think so because I remembered some of the questions having Angoffed them and remember being surprised that so many people got them right or so many people got them wrong – I think I had that stored away in a part of my brain that could have been used for something
more useful, but, nonetheless, I did remember some of those questions, so I think that did
influence me a little bit with the order that we did them... and had we done it, had I been in the
other group it might have influenced me you know on the flip side.

G2-8: I was in group 2 so I agree for me, I think the item mapping may have influenced my
Angoff that I just, like we talked about those questions a lot in the item mapping session and then
when I did Angoffing I recognized the questions and I tried to kind of put it aside but it was
oftentimes hard to.

G2-1: So I was in group 2 and I do recall the controversial items, and, what that alerted me to
was that I needed to consider, I think it did help me in that I needed to consider that perhaps I
was overestimating or underestimating the candidates or the minimum-qualified, and, but it
wasn’t a bad thing, it almost gave you a sense of ‘ok, this is one item I need to consider a little
more carefully’ so it almost complemented the Angoffing.

Facilitator: I will move forward and agree feel free to stop me at any time. So here’s how the
next part of our conversation will go: I wanted to start with talking about those methods overall
and what did you think of them, so thank you so much for that. Next we are going to transition to
the questions I asked you in the most recent survey, and thank you all again, for answering. The
first four questions were about maximum acceptable passing score, minimum acceptable passing
score, so I want to go through those really quickly. The first question was ‘What do you think is
the maximum acceptable passing score (or cut score, percent needed to pass) on this module?’
My first question for you all is do you understand the question being asked here?

G2-6: Sort of.

Facilitator: Can you elaborate?

G2-6: I had to think about it for a long time, the way it was worded seemed odd to me, I think I
understood the concept of setting a floor- because I understand it. It’s not really a maximum,
that’s what I found kind of confusing. I guess we’re trying to just set a range as to, within this,
what we think an acceptable lowest passing score might be and that’s kind of how I approached
it, with the maximum score and minimum score might be.

G1-2: I was glad you gave an example because I found the wording a little confusing, I thought I
would call that a minimum passing score, I’m sure that’s not correct in your world, but the
terminology was a little confusing to me. I think I did understand the question by the time I
answered it.

G2-8: I don’t think I understood the question.

G1-1: I think I misunderstood it too – go ahead.
G2-8: The second question was what’s the minimum acceptable passing score, if I remember correctly, so then I was confused, like ‘well it can’t be the minimum passing score,’ so I thought what was the highest possible score that would be acceptable. So I may have misconstrued the question.

G1-1: I did the same, I understood the question that way as well, so I thought it was acceptable to get everything right, so I’m one of the 100%-ers.

Facilitator: I understand that this question might be confusing – that’s very helpful feedback for me. So for the next question – same thing here - what is the minimum acceptable passing score, we started to talk about this as well I think some of you said the same thing, this was kind of confusing as was the question previously, is that correct, or is this one more understandable do you think? What do you think about the data we got here?

G1-1: I thought this was a question that made logical sense to me, and I understood the question and answered it, what I think the minimal acceptable passing score would be on the modules, very clear.

G1-3: I thought you meant the question to be what would the test be too easy and what would the test be too hard if we set the number too high or too low as a pass.

G2-6: Right I think it’s the same way I thought of it, what I was saying before, kind of a range of acceptable passing scores would have to fall into that.

G1-3: So if you make the score 100, that’s too high because a lot of people are going to fail. If you make it 60, that’s too low, because a lot of people are going to pass who don’t deserve to pass.

Facilitator: Did you understand that, some of those of you who haven’t spoken up yet?

Panelist unknown: Yes

G1-2: Yes, I understood it the same way G1-3 did.

Facilitator: Given your knowledge of the qualifications and performance of the examinees that take this module, what do you think is the maximum failure rate that would be acceptable? So a range here from 2% to 25% [later clarified range to be 2 to 50], so same question to you all – thoughts about this question, did you understand it, and what did you think about in arriving at your answer here?

G1-3: This one is harder for me because we know the pass rate for the examinees, the residents, we fail a heck of a lot more than in MOC, so it depends a little bit. We accept a 25% fail rate on the residents’ side, but only a 2% or 5% on the [exam name] side.
G2-6: Like G1-3 I thought of this as practical terms of: what are we trying to achieve with the [exam name]? And I think the idea is that most people should pass.

G1-1: That’s always my thinking with the [exam name] exam, that everyone who is taking this examination has passed the written examination and the oral examination, and it’s a recertification test, so we would hope that the vast majority of them had the knowledge to continue to be a certified ophthalmologist. So it seems to me that a low percentage that would fail the test, that’s always the way I approach this type of question.

G1-3: A 40% fail rate on the [exam name] would be a dramatic negative and profound indictment of our medical system.

G1-6: Yes I agree.

Panelist unknown: I still go back to, not what the failure rate should be, but [inaudible] what is important? And of course we want to ask questions that we think are necessary information and, so I still think of the questions not necessarily a failure rate, but I think a high failure rate would be terrible and when you look at the population of people that you work with [inaudible] wouldn’t think that 40% should fail, but it still goes back, for me, to the material.”

Facilitator: I’m going to move on to the last of these four questions that are kind of worded like this which is ‘what do you think is the minimum failure rate that would be acceptable, and this is one where there is a little bit more - I guess I should say a little less variance in the answers. Is it acceptable if 5% of individuals fail? 0%? There’s only one rating of 0 here, so I’ll pose that question to you again. Given what you know about the population of individuals that take the [exam name] core, what do you think is the minimum failure rate that would be acceptable? Is it acceptable if no one fails?

G1-3: It would not pass the public sniff test if no one failed.

G2-6: Right.

G1-2: Agreed.

G1-8: I agree.

Panelist unknown: In fact I would propose it wouldn’t pass our sniff test.

G1-4: I agree too, but I also think we are not doing this on a, some sort of fail, as far as, after the fact. Am I right, Facilitator? Am I right about that? This many people have to fail, this many people have to pass…

Facilitator: Correct. You’re right, so these questions were a little challenging because they’re kind of asking about, you know, what percentage of people you think is maximum, minimum,
but you’re right. When we set a criterion-referenced standard, it is always possible for everyone
to pass or everyone to fail. At the end of the day, the standard is going to be set like: you have to
get x out of 50 questions right, and it’s possible for everyone to pass. This is more theoretical –
knowing what you know about your colleagues, the diplomates, the whole population of people
taking this assessment, do you think it’s acceptable if there’s a zero percent pass rate? Do you
think it’s acceptable if it’s, you know, 40 percent? Now maybe that’s too high, so I’m trying to
get kind of that idea of this range. What makes sense as far as a failure rate? What would you see
and say, “Oh yeah, that makes sense to me.”?

G1-1: So if I’m going to use an analogy here, I just went to our medical school student progress
committee. We look each year at the USMLE step 1, step 2 pass rates for our class. Now
obviously nationally, it’s set so that a certain number of students fail at the national level. So if
you look at our medical school, we have a handful that fail every year and those are the ones we
need to focus on. But once in a blue moon, there was one year where everyone passed. And it
was great! And it was just one of those years and that’s why on this question I felt like, if one
year everybody studied really hard and no one failed, it would be ok. So for a minimum failure
rate, zero would be ok.

G1-4: Right, I agree. It’s hard to say that board certified folks and they’re in that group that are
having to recertify that we don’t have confidence in them. Not to say that it’s not surprising if
some would fail, but we shouldn’t have the goal that some should fail.

Facilitator: This is really a judgment – I’m curious what you all think of the population in
general and what your expectations are about pass and failure rates on this exam and these
numbers are not going to impact what the cut score is going to be, nor will I say “well I know it’s
not a zero percent failure rate is ok, so let’s not fail anybody.” This is really just to get a sense of
what do you think is acceptable? What do you think is normal? Would it surprise you if no one
failed? And I’m hearing a few say no. I mean that’s possible. And maybe it would surprise me if
nobody fails, so I’m getting that idea from you all that the number would be small, which is what
we are seeing on this slide. Any additional thoughts on that?

G2-1: Yeah so, the questions when I was going through them and re-thinking what you were
asking, I got the sense that you were wondering about our philosophy about what this test
accomplishes, what its goals are. So I’m actually a little in flux about that because as I write
items I have different populations in mind, and they range from, on the high end, where someone
is likely to fail because they’re on the high end, and by that I mean there are physicians in
practice actively doing major things and are so subspecialized, for example, in diabetic
retinopathy only, that that comprises 100% of their effort, and they’re doing a lot of productive
things, and yet when they have to take something like this, they haven’t seen glaucoma in ages.
They’re aware of it, but it’s just, that is a population where they are on the highest end of
subspecialization and really good doctors and then we have to ask them to study not only the rest
of retina, which they don’t see because they are so subspecialized, and then the rest of core. So I
keep that in mind sometimes, and then the population where they’re ten years out is going to be
different than the population that’s 20 years out and subspecialized. So there’s a whole range of
individuals that are productive, up to date, in their field, but are going to be – but to prepare for
the [exam name] is quite a time-consuming effort, so that’s why I tended to think well maybe the
exercise in preparing for [exam name] is to get this question right, get this question wrong and
then score it. But the whole exercise of preparing and maybe there should be more of a
forgiveness in some populations because it really is pretty far from what they’ve been doing for a
while.

Facilitator: Right. Very interesting comment, yeah.

G1-4: It may not be any kind of reflection of what kind of physician they may be because some
of the questions are certainly more niche.

Facilitator: Right. And what you all are getting at is, is this test a perfect snapshot of somebody’s
capability in their current practice setting? And the answer is of course “no.” It’s not. It is the
best surrogate we have for measuring whether someone has some baseline level of competence
as necessary to maintain a certificate which, to be honest, is just in “ophthalmology.” We really
appreciate that perspective. It is something the board talks about a lot and something we
internally discuss as far as measurement and subspecialization. Slightly outside of the scope of
where we might be going next in our discussion for tonight, but I really appreciate that comment,
but it is something I assure you the board talks about and is trying to keep on the pulse of as we
try to better our measurement tools and get more and more down that road of being able to offer
a customized exam experience for individuals.

Facilitator: so now the final part of our discussion is going to be about the actual cut scores. So
we talked about the Angoff method. We talked about item mapping. I asked you some weird
questions about what results you’d expect on this test overall. Now let’s talk about what actually
happened. You all submitted Angoff ratings, and I compiled those. If we used those to set the cut
sore for this examination, both groups actually arrive at the exact same cut score. So I would
have presented group 1’s cut score then group 2’s cut score, but they were the same: 39 out of 50
questions would need to be answered correctly in order to pass. If that cut score were
implemented, using the data I have from last year, about 88% of examinees would pass, and 12%
would fail the module. And then, if we used the cut score from item mapping – so group 1 set a
cut score of 66% correct or 33 questions out of 50. If that cut score was implemented, 99.5% of
examinees would pass, and approximately .5% would fail. Group 2 – that’s our group that did
the item mapping first before Angoff ratings, set a cut score of 56% correct or 28 questions out
of 50. If that cut score was implemented 99.9% of examinees would pass and .1% would fail. So
I’m going to ask you about what your response is to these cut scores if you think they’re
reasonable or not. I asked you that in the survey and I’m going to bring that up. (Survey
responses to this question were shown). First let’s talk about Angoff, that would result in an 88%
pass rate. Does that seem acceptable to you? What does 39 out of 50 seem to you – too high?
Too low? Acceptable?
G2-8: I thought it was a little too high – I said disagree. I mean it was nice that both groups were consistent, and we agreed. But, I think that pass rate [means cut score] is maybe a little bit too high.

G1-3: I was surprised that that many people would fail given that cut score.…

Facilitator: By that you mean you assumed more people would answer 39 questions correctly?

G1-3: Yes. The cut is reasonable, but the outcome of that cut is scary. Too high a fail rate – but that number seemed right. Can you tell us if one or two questions one way or the other would make it better?

Facilitator: I don’t have that data in front of me, but I will say this is a meaty part of the curve, and by that I mean a lot of people are scoring around this, a little higher than this. So going up or down the scale a couple of points would make a difference, yeah.

G1-3: So from my perspective this number seems right, but if it was 75 [percent correct as the passing standard] and that resulted in a 5% fail, I’d be so happy with that. I think that’s fair. This Angoff number seems right to me, but that fail rate bothers me.

G1-4: It also makes me wonder how well we’re doing with our questions, you know, I mean, if you know you can get eleven questions wrong out of 50, you have to get more than that to fail, and that many people are getting that many wrong. Like you said, you’re surprised – that I’m wondering how fair our questions are… because I just, again these are people who have been board-certified and passed and did all that work and passed the orals and just recertifying. I also wonder, Facilitator, are there people who are out 30 years (I mean, we don’t have this yet), people who are out 30 years versus 10 years, or I guess we could compare 10 years versus 20 years now. Is there a difference in their pass rate?

Facilitator: I don’t know that offhand, I’m sorry I haven’t looked at that data. So I don’t know. I just know overall.

G1-3: To me it doesn’t pass my own personal knowledge base list of my peers. I would not say one of eight of my peers would fail.

G1-4: Right. So I feel like it’s not reflecting what we know about our colleagues and how they practice. Again it’s getting to… maybe the core is something we shouldn’t be doing.

G1-7: I think if you go back to the minimum maximum question, and, if I remember the minimum pass rate that kind of fell in the middle of the numbers was like 70, 75 – and, if you look at that, perhaps it falls more in line of just low, like maybe 37 out of 50 questions, and then that would give like 5% not passing.

Facilitator: Is there anyone that said agree or strongly disagree? Any thoughts?
G1-2: I’m wondering for the you know occasionally we’ll get a question where a lot of people get it wrong and looking at it you can see where they perhaps were misled or maybe it wasn’t a great question. There really isn’t a way to correct for that [inaudible] sometimes throw questions out?

Facilitator: Yeah. Immediately after the administration if we find that a question has unusual statistics, and by that I mean less than 60% of people answered it correctly or it has a negative biserial or anything else that just raises a red flag, we send it back for review. And, if we find it’s flawed, like ‘yeah a and c could be an answer,’ then we’ll toss it out and either give everyone credit for it or give everyone that answered a and c credit because we think both of those are possible answers. So we do kind of correct for that problem. The people that scored 39 out of 50 got 39 out of 50 good questions wrong [right]. We’ve already corrected for anything we didn’t think was fair to score on.

G1-2: I’m wondering for the you know occasionally we’ll get a question where a lot of people get it wrong and looking at it you can see where they perhaps were misled or maybe it wasn’t a great question. There really isn’t a way to correct for that [inaudible] sometimes throw questions out?

Facilitator: Ok let’s move on to item mapping. Group 1, you arrived at a cut score of 33 out of 50 items, and I asked, ‘Do you think this cut score is acceptable?’ These are so interesting, these results, because you’re divided among the agree and disagree categories. Again, that standard would result in a 99.5% pass rate. What are your thoughts on this potential standard?

G1-6: Seemed like too high of a pass rate for that many questions missed.

Facilitator: You think the standard should be set a little higher? Or you’re surprised what the pass rate is corresponding to this standard?

G1-6: You said 99.5% of examinees would pass by getting 33 questions correct, correct?

Facilitator: Right.

G1-6: I don’t know. It seems like less should pass from missing that many questions in my opinion.

Facilitator: Do others agree or disagree? Do you think this standard is too easy or too low?

G1-1: Well, I, looking at the distribution of questions and how many individuals got some right, part of my thinking was, in order to have a distribution on the test, you have to have some questions that are more challenging than others, which is why we have that spread, and the questions are obviously challenging even to someone who is board certified in that practice and that might reflect, to some degree, some of the specializations that we were talking about where we, even though we consider this, we write questions to the core, some ophthalmologists no longer practice core ophthalmology, so, even if they studied, some of these concepts may be difficult, so I think it’s acceptable to miss that many, and, obviously, you would hope that folks do not miss that many, but I thought it was acceptable if they missed that many, um, hopefully they would not miss that many, but it was a little, I thought the cut score could be a little bit…
you said 99.5% with this method? Facilitator, is that right? Yeah, that seems a little high. I may have been in the disagree group but not sure.

Facilitator: Anyone else on this one?

Panelist unknown: I think it would be too easy too – I agree with what everyone else has said about that. Well, most people - the majority I think.

Facilitator: Does anyone strongly agree that this is the standard that we should recommend? [no response]

G2-1: Yes. In doing the four questions in advance, it gave me an idea of where I felt a range should be, and so then, when juxtaposed to these actual percent interpretations of the item mapping method and so forth, it just seems that then you have a real reaction to it based on the fact that you had considered the four questions in advance. So that’s what I’m feeling is conflicting, whereas the real numbers are showing one thing, my feeling in answering the general range was different.

Facilitator: Any other thoughts about this standard? [no response] Alright let’s move on to our last standard, which is the group 2 set standard, which was a 28 out of 50, which would result in a 99.9% pass rate, a .1% failure rate and so again, I’ll ask the question, is that a reasonable standard?

G2-1: It just seems like a really low bar to only have to get about half the questions right assuming the questions were well thought out and really represent what ophthalmologists should know.

G1-6: I agree

G1-4: I agree too. I just think, I keep going back to are we missing something if our questions … somebody can get that many wrong and pass as far as our questions. There’s a question about how a tissue from a lid lesion has to be processed. A retina specialist who has been out there for 15, 20 years is not going to know that unless they take a [exam name] preparation course, and, I wonder how many questions we have that are like that.

Facilitator: Yeah. You all did see all the questions. I certainly don’t expect you to have memorized them, by any means. Given that this is the test, let’s say these are the questions and these will be used forever, hypothetically. Do you think the standard should be a little bit lower to reflect that, or do you think they were pretty acceptable for everyone to know and the standard should be a little higher?

G2-1: I almost feel like there could be different standards for different populations taking it, like you were implying there’s a customized exam, having the 90% pass for the person that’s 25 years out only doing one thing well is ok, and the expectation would be different for someone
who is doing general and is ten years out. So it’s almost like we are trying to set one standard for a really wide range of individuals taking this test.

Facilitator: Think about the individual that needs to know the least and what performance level would be acceptable.

G1-4: Facilitator, is the core weighted as heavily as the other two modules? Say you’re a retina person and you choose to do two retina and obviously the core, they are weighted equally?

Facilitator: Yes. Each question counts for the same amount on weight. Two-thirds would be on retina and one-third on core.

G1-4: I just wonder if we need to rethink our core questions, think of it more like what could walk into your office as any sort of specialist, where to go next with it, not necessary what the technique would be in the O.R. or what the tissue sample would need to be placed in or what the options are for treatment but maybe more differential… I don’t know [inaudible] as far as how many questions could get wrong and pass in this one module?

G1-8: I think, there’s only one core they’re going to be required to take, and the other is going to be their specialty. And looking at this test overall and all of the questions, there were probably a handful that I felt like, ugh, maybe that is not an appropriate core question, but I didn’t think it was an overwhelming number of questions. And so, I mean, I feel like it wasn’t you know, unacceptably difficult in terms of the knowledge that we’re requiring, but maybe I’m just a really tough person.

G1-4: I don’t know the answer, I’m just kind of throwing it out there, but I’m also not an oculoplastics surgeon who hasn’t necessarily looked at, used a slit-lamp for, I don’t know [inaudible] and, you know, know significant anterior segment necessarily, or retina, I don’t know.

G1-1: I think that’s what we’re discussing here, it’s very difficult to design a test um, that includes core knowledge that all ophthalmologists should know I think that’s what we are finding with these questions, that they, some of them, it is very difficult, because what an oculoplastics surgeon needs to know versus a diabetic retina specialist needs to know are so different that it’s hard to find key questions that both of them should be able to answer. I think that says, I think we’re heading back to, just the design of the test is difficult, which is why we have some questions on there like pathology [inaudible] which may not apply to everyone so, it is difficult, that doesn’t answer Facilitator’s question of the night – this is philosophically a difficult test to design.

G1-4: In reality, practicing, even if they knew that, I can’t just throw this tissue in anything. I need to call the pathologist and ask them what they would like it in. That’s acceptable out in practice, but that doesn’t, that can’t come up in these sorts of tests, so um, it’s hard to reflect, in a test like this that somebody um, not qualified to practice and continue practicing, and I’m also
considering that, those who - I don’t know if those statistics are available- those that do not pass
the [exam name], if their other aspects of their maintenance of certification and don’t look so
good either like as far as their chart reviews and things like that.

Facilitator: I have some final thoughts [questions]. Do you think the order of the activities
affected your ratings? We already talked about that a bit. Which method did you feel you were
able to apply more reliably related to making yes/no ratings versus 0-100?

G2-8: We talked about this a little bit already, but I think the Angoff method may have been a
little more reliable just because you could think about it on your own and not be influenced by
anyone else. I mean I liked the item mapping method, but I think, for me, the Angoff method for
me might be a little bit better.

G2-1: I felt that the item map, I did the item mapping first and the Angoff second, and the item
mapping was almost like you got to preview and bounce – get some feedback outside of yourself
and so it was a good preview and set the stage for the Angoffing. So again I said I think it
compliments it – the Angoffing.

G1-4: And I did the opposite, and I didn’t really think it had a lot of influence the flip way. I
don’t think my Angoffing first influenced my item mapping.

G2-1: The item mapping made you round up or down, because it was a yes/no, whereas the
Angoffing you could put in a little bit of nuance. I don’t know if that affected the final outcome?
But it really – the item mapping you just had to – it was a yes/no, you had to round up – it was
very binary, whereas the other had a little more range.

Facilitator: Did any of you find the cognitive complexity of coming up with a number to be more
or less challenging than coming up with a yes/no?

G1-4: Not really. In some ways coming up with a yes/no is a little more difficult.

G2-8: I thought the same for both, cognitively speaking.

G1-1: I felt the binary concept was more difficult – I think, again, because it was a group
activity, I thought it was more difficult to come to a conclusion versus the Angoff where
individually you’re scoring it seemed easier to come up with a comfortable score.

G2-1: The other thing is that with the item mapping, if you disagreed with the yes/no and you
were the minority, you’re - you had to accept the majority, and so it’s as if you didn’t have the
liberty that you do in Angoffing to decide for yourself and stick to it or not stick to it. So there’s,
there’s a little bit of that. That happened, actually, not too infrequently.

G1-8: I preferred coming up with a number because I felt like there was more precision, I think
somebody mentioned you could more “fine tune” your response. There was more nuance to it.
Facilitator: Thank you. I’d just like to open it up - any final thoughts on the methods? Moving forward is there one that you’d recommend?

G1-3: Angoff

G2-1: I vote for Angoff, because of the-

G1-4: Our results of the Angoff were a little frightening!

Facilitator: Which method – you didn’t like the Angoff results, but you liked that process better?

G1-4: I did – that’s what I’m saying, I found the process a little, especially when we got the feedback, “Well here’s what you put,” so we basically did it twice, but, yeah.

G1-1: I also, I prefer the Angoff method and I think that, especially knowing what Facilitator told us which is that, even, I understand – it is scary to think about the cut scores we created with Angoff and perhaps they were off a little bit, but you realize we did the first part and typically there’s going to be another review before any decision is made. So I’m still comfortable with it. I think it is, like G1-8 said, I like the degree of precision for each individual question you’re developing a score, which I like.

G1-4: Yes.

Facilitator: Anyone else have a preference between the two methods you want to suggest moving forward?

G1-8: Angoff.

[Question sent through chat box via the webinar by G2-5]: Facilitator, I thought the Item Mapping worked well, but I thought at the time that we were establishing a low (easy) pass threshold because of time. I was in the group that did item mapping first. I think if we had gone through a few more questions, we might have landed on a tougher (and therefore more widely accepted) cut rate. It may be worth trying again another year, particularly with folks who are more experienced with both methods.
Curriculum Vitae

Sarah D. Schnabel
Sarah.d.Schnabel@gmail.com, fowle@uic.edu

Education

University of Illinois at Chicago
PhD Candidate, Educational Psychology, 2012 – 2018 (expected)
4.0/4.0 GPA

4.0/4.0 GPA

Advisor: Everett Smith, PhD

University of Illinois at Urbana-Champaign
Bachelor of Science (BS), Psychology, 2003 – 2006
Minor in Spanish
3.78/4.0 GPA

Teaching Experience

Large Scale Testing (EPSY 562), University of Illinois-Chicago, fall 2016
Primary instructor, online course (overall evaluation 4.3/5.0)

Presentations


Presentations (continued)


Professional Service

- American Board of Medical Specialties Research and Evaluation Collaborative – Member (2015-present)
- American Board of Medical Specialties Longitudinal Assessment Task Force – Member (2016)

Employment

American Board of Ophthalmology Psychometrician
January 2014 – Present

National Association of Boards of Pharmacy
April 2008 – January 2014

Competency Assessment Program Development Manager
October 2012 – January 2014

Graduate-Level Coursework in Educational Psychology

- Computer Adaptive Testing
- Rating Scale Analysis
- Advanced Analysis of Variance
- Multiple Regression
- Large-Scale Assessment
- Data and Interpretation in Educational Inquiry
- Educational Measurement
- PhD Proseminar I and II
- Research Design
- Item-Response Theory
- Analyzing Rating Data
- Independent Study
Graduate-Level Coursework in Educational Psychology (continued)

- Assessment for Measurement Professionals
- Philosophical Foundations of Educational Inquiry
- Theories of Educational Psychology
- Multivariate Analysis of Data
- Achievement Motivation

Software Knowledge

Proficient in Microsoft office suite and SPSS. Basic working knowledge in SAS, Bilog-MG, WINSTEPS, FACETS, and R.

Memberships

American Educational Research Association (AERA)
National Council on Measurement in Education (NCME)