2014

EFFECTS OF ITEM-LEVEL FEEDBACK ON THE RATINGS PROVIDED BY JUDGES IN A MODIFIED-ANGOFF STANDARD SETTING STUDY

Michael R. Peabody
University of Kentucky, michael.peabody77@gmail.com

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Michael R. Peabody, Student

Dr. Kelly D. Bradley, Major Professor

Dr. Robert Shapiro, Director of Graduate Studies

EFFECTS OF ITEM-LEVEL FEEDBACK ON THE RATINGS PROVIDED BY JUDGES IN A MODIFIED-ANGOFF STANDARD SETTING STUDY

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Education
at the University of Kentucky

By
Michael R. Peabody

Lexington, KY

Director: Dr. Kelly D. Bradley, Professor of Educational Policy Studies and Evaluation

Lexington, KY

2014

ABSTRACT OF DISSERTATION

EFFECTS OF ITEM-LEVEL FEEDBACK ON THE RATINGS PROVIDED BY JUDGES IN A MODIFIED-ANGOFF STANDARD SETTING STUDY

Setting performance standards is a judgmental process involving human opinions and values as well as technical and empirical considerations and although all cut score decisions are by nature arbitrary, they should not be capricious. Establishing a minimum passing standard is the technical expression of a policy decision and the information gained through standard setting studies inform these policy decisions. To this end, it is necessary to conduct robust examinations of methods and techniques commonly applied to standard setting studies in order to better understand issues that may influence policy decisions.

The modified-Angoff method remains one of the most popular methods for setting performance standards in testing and assessment. With this method, is common practice to provide content experts with feedback regarding the item difficulties; however, it is unclear how this feedback affects the ratings and recommendations of content experts. Recent research seems to indicate mixed results, noting that the feedback given to raters may or may not alter their judgments depending on the type of data provided, when the data was provided, and how raters collaborated within groups and between groups. This research seeks to examine issues related to the effects of item-level feedback on the judgment of raters.

The results suggest that the most important factor related to item-level feedback is whether or not a Subject Matter Expert (SME) was able to correctly answer a question. If so, then the SMEs tended to rely on their own inherent sense of item difficulty rather than the data provided, in spite of empirical evidence to the contrary. The results of this research may hold implications for how standard setting studies are conducted with regard to the difficulty and ordering of items, the ability level of content experts invited to participate in these studies, and the types of feedback provided.

KEYWORDS: Standard Setting, Angoff method, Rasch model, Rater bias, Form difficulty

Michael R. Peabody
Student's Signature

10/21/2014
Date

EFFECTS OF ITEM-LEVEL FEEDBACK ON THE RATINGS PROVIDED BY JUDGES IN A MODIFIED-ANGOFF STANDARD SETTING STUDY

By

Michael R. Peabody

Dr. Kelly D. Bradley
Director of Dissertation

Dr. Robert Shapiro
Director of Graduate Studies

10/21/2014
Date

ACKNOWLEDGEMENTS

No work of this magnitude is completed in a vacuum; there are a great number of people who have contributed in some way, both large and small. I would first like to thank my dissertation chair Dr. Kelly Bradley. Not only was her guidance through the dissertation process invaluable, but she was also the one who introduced me to Rasch measurement and encouraged me to pursue a career in psychometrics. I doubt I would be in the position I am today without the good fortune to have met her when I did. I would also like to thank the rest of my committee: Dr. Neal Hutchens, Dr. Wayne Lewis, Dr. Hongwei Yang, and Dr. Jeff Osborn. Each provided valuable insight and a unique perspective that challenged met to explore new ideas, which ultimately helped to improve the quality of this final product.

I would also like to thank the American Board of Family Medicine for supporting me in this research. Furthermore, I would like to offer a special thanks to Dr. Thomas O'Neill, whose mentorship and guidance has been vital to my completing this project and whose passion for measurement and psychometrics is infectious.

There are also several others who have played some part in my reaching this accomplishment. I would like to thank, in alphabetical order: Kate Akers, Carol Behr, Justin Blackburn, Bruce Holle, Nikki Knutson, Eric Reed, Ken Royal, Jen Soltis, and Baron Wolf.

Finally, I would like to thank my wife Amy and my three children: Dean, Ryan, and Julia. Their love and support made this entire adventure possible. Through any number of long days and nights I always knew that I could count on them for encouragement and motivation. My wife endured far too many nights of single-parenting while I was engaged in some scholarly activity. None of this would have been possible without her support – this accomplishment is as much hers as it is mine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

Introduction

Setting performance standards is a judgmental process involving human opinions and values as well as technical and empirical considerations. Although all cut score decisions are by nature arbitrary, they should not be capricious (AERA, APA, & NCME, 2009; Cizek, 2012; Shepard, 1979). Establishing a minimum passing standard is the technical expression of a policy decision. The information gained through standard setting studies informs these policy decisions. To this end, it is necessary to conduct robust examinations of standard setting studies in order to understand how the information gained from standard setting studies influences policy decisions.

Purpose and objectives of the study

Examining how information regarding item-level feedback influences the perceptions of item difficulty held by content experts is a subject that has not been studied extensively. However, the way in which information regarding item-level feedback influences content experts' decisions may hold extensive consequences with regard to setting an appropriate passing standard and the subsequent pass/fail or other categorical decisions. In particular, Hambleton, Pitoniak, and Copella (2012) call for more research on the results related to the utilization of performance data as feedback in standard setting, specifically noting the interesting questions raised by Clauser, Mee, Baldwin, Margolis, and Dillon (2009) when they provided incorrect feedback to judges.

The current study seeks to add to this body of literature by examining how the item-level feedback provided to content experts affects the ratings they provide.

Research questions for each study

The primary research question guiding this study is, "how does the item-level feedback provided to content experts influence the ratings they provide?" In order to examine this question, I will analyze three different standard setting exercises that were each constructed with this research question in mind. The research questions guiding each of these three studies are:

1. Does the difficulty of the standard setting form affect the ratings provided by content experts?
2. Does the ability level of content experts affect the ratings they provide?
3. Does altering the feedback given to content experts affect the ratings they provide?

Overview of study

As previously noted, the primary research question for this study is, "how does item-level feedback provided to content experts influence the ratings they provide?" In order to fully explore this question, I will conduct three distinct studies designed to examine the various ways in which the final ratings of judges are influenced by the feedback provided, as well as whether feedback has an undue influence on judges based on demographic characteristics.

For the first study, I will investigate whether the difficulty of the standard setting form affect the ratings provided by content experts. Judges were given one of two forms, one with a mean item difficulty scaled score of 200 or another that was targeted to the exam's passing standard with a mean item difficulty scaled score of 390. For the second study, I will investigate whether the ability level of content experts affect the ratings they provide. This study included two cohorts of judges: those who scored 600 or above on the last examination and those who just met the passing standard with a score of 390 or 400. For the third study, I will investigate whether providing erroneous item-level feedback affects the ratings the judges provide. This study provided judges with an inverted conditional p-value of item difficulty as feedback. So, an item that 10 percent of minimally qualified candidates would be expected to get correct was said to have been answered correctly by 90 percent of minimally qualified candidates.

Each of these research studies was conducted as part of an operational standard setting exercise. The examinations associated with these standard setting exercises all reported scaled scored ranging from 200 to 800. Because these were all operational standard setting exercises designed to produce a passing standard for an upcoming high-stakes examination, they were conducted in an extremely uniform and regimented manner. The methods for participant recruitment and training were virtually identical for each of the three studies under examination here. Furthermore, this dissertation is written using a multiple-manuscript format such that each chapter should be able to stand on its own without relying on information provided in previous chapters. Therefore, the tone and tenor of the document, if each chapter is read consecutively, can appear rather

repetitive. This is not so much a flaw, but rather a feature; the document was intentionally written in this manner.

Another notable aspect of this study is that the judges provided ratings online and not in-person as is customary in the vast majority of modified-Angoff standard setting exercises. The asynchronous process of providing ratings allows for the elimination of any group-effect and focuses on the individual judge's perception of item difficulty rather than focusing on consensus building within the group of judges. The goal of each method is the same, to produce a minimum passing standard, but the process and reasons for providing the specific types of feedback is slightly different. The asynchronous process first asks the judge to answer the question; whereas the in-person method does not typically ask this. The goal of this feedback is to assist the judge in formulating their perception of item difficulty. If a judge were to answer a seemingly easy question incorrectly, perhaps that item is more difficulty than the judge initially thought. The conditional p-value feedback is provided in order to allow the judge to compare their conceptualization of a minimally qualified candidate to the current minimally qualified standard. This is all done with the intent of assisting an individual judge to formulate the ability of a minimally qualified candidate, whereas the intent of the feedback and discussion typical of the in-person method is to allow the group to come to a consensus around their combined idea of the ability of a minimally qualified candidate.

Limitations

The primary limitation of this study is that it is correlational research and, while useful to help uncover the relationship between variables, does not provide and

conclusive evidence for causation and often leads to more questions than answers. There were also several limitations with the design of the studies. As previously mentioned, each of these studies was conducted as part of an operational standard setting for an upcoming high-stakes examination. Therefore, a certain protocol was required to be followed. For security reasons it was not possible to repeat either questions or judges for any of the studies. When examining whether the difficulty of the form of the exam affected the ratings, the ability to overlap items between both forms would have strengthened the research design. Similarly, having some judges rate both forms would have added to the strength of the design, but the additional cognitive load and exposure of items was not an acceptable proposition.

Furthermore, the study in which judges were provided erroneous feedback was accidental and not a planned experiment. A much stronger research design would have allowed for a control and experimental group, both receiving the same items but with different feedback. However, simply because it was not a planned experiment does not mean we cannot learn from our mistakes. It is not uncommon in an operational setting that interesting research questions occur by chance and not by design and researchers should be willing to embrace these unforeseen opportunities.

Finally, at issue is whether the results of this study are generalizable to standard setting exercises in other fields since the judges in each of these standard setting studies were board certified physicians and the effect of different types of feedback mechanisms may affect these highly trained content experts differently than content experts in other fields.

Significance and contribution to the field

The modified-Angoff method remains one of the most popular methods for setting performance standards in testing and assessment (Hurtz & Auerbach, 2003; Plake & Cizek, 2012). With this method, it is common practice to provide content experts with feedback regarding the empirical item difficulties; however, it is unclear how this feedback affects the ratings and recommendations of content experts. Recent research seems to indicate mixed results, noting that the feedback given to judges may or may not alter their ratings depending on the type of data provided, when the data was provided, and how judges collaborated within groups and between groups. The research proposed here seeks to examine issues related to the effects of item-level feedback on the ratings provided by judges. The results of this research may hold implications for how standard setting studies are conducted with regard to the difficulty and ordering of items, the ability level of content experts invited to participate in these studies, and the type of feedback that is provided to judges. In high-stakes testing, setting performance standards is of critical importance and it is imperative that the utmost care be taken to ensure that standard setting exercises are conducted with the strongest theoretical and empirical foundation possible.

Chapter Two

Review of the Literature

Introduction

The research presented here investigates the effects of item-level feedback on the ratings provided by content experts during modified-Angoff standard setting exercises. These particular standard setting exercises were conducted by a medical certification board for several high-stakes certification examinations. High-stakes testing in the medical profession has a long history and physicians are tested early and often throughout their professional development.

The first medical certification board was the American Board for Ophthalmic Examinations (now the American Board of Ophthalmology), which held its first examination the University of Tennessee at Memphis in 1916 (Cordes & Rucker, 1961; Shaffer, 1991). The second medical certification board, the American Board of Otolaryngology, was incorporated in 1924 and the American Board of Obstetrics and Gynecology became the third board in 1930. By 1960 there were 19 specialty boards and there are currently 24 member boards of the American Board of Medical Specialties (ABMS) (American Board of Medical Specialties, 2014).

Young, Chaudhry, Rhyne, and Dugan (2010) conducted a census of licensed physicians in the 70 state and territorial medical and osteopathic boards and determined that there are 633,733 licensed physicians that are certified by an ABMS member board. This number represents 74.5% of all physicians licensed to practice in the United States.

Increasingly, hospitals that employ physicians and insurance companies that provide malpractice coverage are requiring board certification. The consequences for either not enrolling in a certification program or losing board certification can range from being unable to find employment at local hospitals to losing malpractice insurance. Although board certification is optional, the financial ramifications for not participating in a certification program can be rather substantial.

Beginning with their application to medical schools, prospective physicians take the Medical College Admission Test (MCAT). Once admitted to medical school, students take a four-part series of United States Medical Licensing Exams (USMLE): Step 1 after their second year; Step 2 CK and Step 2CS during their fourth year; and Step 3 at the end of their first year of residency. Following medical school, most physicians enter into a residency program for their chosen specialty, the duration of which varies by specialty. Typically, the sponsoring board for a residency program will provide an annual in-training examination to residents, which is designed to assist residency programs in assessing the relative strengths and weaknesses of their residents. At the completion of residency, physicians may choose to sit for initial certification by their sponsoring board. If successful, the newly board-certified physician will enroll in a 10-year cycle of maintenance of certification (MOC) that culminates with another examination at the end of the 10-year cycle. This brief description of the testing regime for medical students, residents, and practicing physicians helps to illustrate the unique relationship that the medical community holds with the testing industry. The view within the medical community that board certification is an aspect of public protection and advocacy further enhances commitment of medical boards to ensure that testing, and

setting the performance standards for those tests, is conducted in a well-researched manner that adheres to best practices. In this light, it should be noted that all of the relevant research on examining the effects of item-level feedback to be reviewed in this chapter was conducted by researchers at the National Board of Medical Examiners.

Introduction to the Angoff method

William Angoff introduced the standard setting method that bears his name in a chapter for a measurement reference book *Educational Measurement* (Thorndike, 1971). Angoff devoted two paragraphs and a footnote towards outlining the method. The focus of his chapter was score scaling, equating, and transformation; his mention of setting a passing score was incidental and due to this lack of specificity, numerous modifications have been made over the years. Although not particularly relevant to this research, it is of interesting historical note that Angoff credited Ledyard Tucker with creating the method, although it seems to have done little good as this method continues to be referred to as the Angoff method.

The Angoff method (Angoff, 1971) is one of the most familiar and often used standard setting method (Hurtz & Auerbach, 2003; Plake & Cizek, 2012). This method asks content experts to determine whether a "minimally acceptable person" could answer specific items correctly. The content experts often represent multiple stakeholder groups ranging from those directly affected by the standard setting outcome to members of the general public (Hambleton & Pitoniak, 2006; Loomis, 2012). Using the original Angoff method, each item is scored yes(1)/no(0) and then the total score is summed to produce a raw score for the "minimally acceptable person" (Angoff, 1971). Adaptations to this

method are often referred to as *modified-Angoff methods*. These modified-Angoff methods often ask content experts to provide the probability of a minimally acceptable person answering a question correctly (Plake & Cizek, 2012). The sum of the probabilities is divided by 100 to produce the percentage of questions a minimally acceptable person should get correct in order to pass the exam.

Often, the modified-Angoff method involves multiple rounds of ratings in which content experts provide their individual ratings in a group setting, are then provided some kind of feedback, and then attempt to come to a consensus as a group for a final rating. Research has shown that several training rounds should be conducted before judges begin the exercise in full (Jaeger, 1989; Livingston & Zieky, 1982; Plake, Melican, & Mills, 1991; Reckase, 2000, 2001; Reid, 1991).

After judges provide their initial ratings, it is customary to provide them with some type of feedback. This feedback is either aggregated information, such as past exam pass rates, or item-level information such as past examinee performance; the percentage of examinees answering the question correctly as each decile level, the percentage of examinees selecting each distractor, overall item difficulty calibrations, and conditional p-values (Plake & Cizek, 2012).

Criticisms of the Angoff method

A primary criticism of the modified-Angoff method has been with regard to judges' ability to make an accurate determination of the probability that a minimally acceptable candidate would get a question correct. It has been found that in the absence

of performance data, judges may misestimate examinee performance (Busch & Jaeger, 1990; Clauser, Swanson, & Harik, 2002; Cross, Impara, Frary, & Jaeger, 1984; Impara & Plake, 1998; Reckase, 2000). Impara and Plake (1998) found that judges would typically underestimate the performance of minimally qualified candidates on items and that the proportion of items being underestimated was greater for difficult items than easy items. However, this finding is at odds with Shepard (1995) and Goodwin (1999). Goodwin (1999) noted that judges were more likely to overestimate the ability of minimally qualified candidates, while Shepard (1995) found that judges were more likely to underestimate the success on easy items and overestimate success on difficult items. Clauser, Mee, et al. (2009) found evidence to support Shepard and Goodwin, while also noting that this discrepancy with Impara and Plake (1998) may be based on using different definitions of minimally qualified or borderline groups.

Shepard, Glaser, Linn, and Bohrnstedt (1993) concluded that the Angoff method was "fundamentally flawed" (p. 132), claiming that panelists were incapable of making consistent and reasonable judgments. However, this claim drew a sharp rebuke from a number of researchers who claimed that the methods and findings of the were not supported by relevant psychometric literature and misrepresented the findings of other studies (Cizek, 1993; Cizek & Bunch, 2007; Hambleton, 2001; Hambleton et al., 2000; Hambleton & Pitoniak, 2006; Loomis & Borque, 2001).

An additional criticism is that the judges are too reliant on the feedback provided (Hurtz & Auerbach, 2003; Maurer & Alexander, 1992; Truxillo, Donahue, & Sulzer, 1996). Clauser et al. (2002) found that ratings were substantially influenced by the

performance data feedback and noted that the group discussion allowed the ratings to converge towards a mean, but until feedback was provided the correlations between ratings and conditional p-values remained at pre-training levels. Busch and Jaeger (1990) found a similar effect with the correlation of ratings with conditional p-values only increasing following the introduction of performance data. This group effect, allowing the judges to come to a group consensus, has generally been thought of as a benefit of the modified-Angoff method (Hambleton & Pitoniak, 2006; Reckase, 2001). However, these studies seem to indicate that the group effect is less important than the feedback provided.

Finally, Stone (2004, 2006), Stone, Beltyukova, and Fox (2008), and Stone, Koskey, and Sondergeld (2011) criticize the lack of judge agreement and further add that the Angoff method lacks the representation of a salient construct and utilizes validity arguments that misrepresent the nature of evaluation. These criticisms have drawn no response and have been largely ignored by the wider standard setting community.

Conditional p-values

The modified-Angoff method utilized here is a content-based method for recommending a passing standard that asks content experts to examine each item and determine the probability of a "minimally competent examinee" answering a question correctly. Judges are commonly provided some form of past examinee performance data to assist in their decision-making process. Each of the examinations under consideration here was scored using the dichotomous Rasch model (Rasch, 1960). The Rasch model is a logistic model of latent traits that provides person measures and item difficulties in log-

odds units, commonly referred to as logits. There are two options available for providing performance data to judges regarding the probability of a minimally competent candidate answering a question correctly: (1) aggregate the item-level responses only from previous examinees who scored at the passing standard, and (2) calculate conditional p-values for each item. Using the responses from previous examinees may provide a more accurate reflection of the ability of minimally competent examinees, but there is typically a dearth of examinees who scored at the passing standard for each item under examination. Therefore, the typical action is to create conditional p-values for each item. Conditional p-values are calculations of the percentage of candidates with ability estimates at the passing standard expected to get the question correct.

Providing judges a calculation of conditional p-values rather than overall item difficulty is done in order to give standard setting judges a more accurate view of how minimally competent examinees would actually perform on this item rather than relying on judges trying to estimate a probability of success for each item based on their own sense of how a minimally competent examinee might perform. However, the overall item difficulty is used to calculate the conditional p-values. The transformation of the overall item difficulty into a conditional p-value was accomplished using the following formula:

$$\Pr\{\beta_{MPS} = 1\} = \frac{e^{MPS-\delta i}}{1 + e^{MPS-\delta i}}$$

Where:

$\Pr\{\beta_{MPS} = 1\}$	is the probability of a correct response by a minimally qualified candidate
MPS	is the calibration of the minimum passing standard

δi	is the difficulty of item i
e	is the base of the natural logarithm

The overall item difficulty calibration in logits from previous exam administrations is subtracted from the calibration in logits for the minimum passing score to produce a conditional difficulty measure of each item for a minimally qualified candidate. The base of the natural logarithm for this new calibration is divided by 1 + the base of the natural logarithm, which produces a conditional probability. This conditional probability is multiplied by 100 in order to return a percentage in a whole number that judges can readily understand. This percentage is referred to as the conditional p-value of an item.

This is a brief description of using the Rasch model to create conditional p-values; however, it is possible to calculate conditional p-values using other item response theory models. Some of the studies to be discussed below (Clauser, Mee, et al., 2009; Clauser, Mee, & Margolis, 2013; Mee, Clauser, & Margolis, 2013) utilize the 2-parameter logistic model (Hambleton, Swaminathan, & Rogers, 1991).

Review of feedback literature

As previously noted, the ability of judges to provide consistently accurate ratings has been extensively studied; however, the literature on examining the effects of feedback on the ratings provided by judges in standard setting exercises is relatively sparse.

Clauser et al. (2002) examined the initial ratings and final ratings of judges in three groups. The authors found that inter-judge correlation of item difficulty increased following feedback and that a substantial group effect existed. The group discussion

allowed the judges within groups to converge towards a mean rating, but the correlation of ratings to conditional p-values remained at pre-training levels until the introduction of item difficulty feedback. The issues relating to the different ratings for each of the three groups may have begun at training with the discussion of "minimally acceptable persons".

Clauser, Harik, et al. (2009) examined the impact of group discussion and examinee performance information on the ratings of experts. They found that discussion increased inter-judge correlations, but not the correlations with empirical item difficulty. After examinee performance data was provided, the correlations with expert ratings increased substantially, suggesting that discussion without data is of little use in standard setting exercises. The performance data used as feedback for this study was the probability of success for each of five scoring groups: 10% above and 10% below the cut score, the top 10%, the bottom 10%, and two marginal groups where the cut score is close to the center of the score distribution.

Clauser, Mee, et al. (2009) conducted two studies in which the empirical item difficulty provided as feedback was manipulated. In the first study they manipulated the conditional p-values for approximately half of the items by randomly increasing or decreasing the conditional p-values by .5, 1.0, or 1.5 standard deviation units. This manipulated data was incorporated into the performance data provided to judges in the form of performance deciles and distractor information. However, the manipulations using standard deviation units resulted in relatively small changes to the conditional p-values. For their second study, the performance profile for half the items was replaced

with the performance profile from another item. They found that judges incorporated the feedback whether it was correct or not and concluded that judges relied on the data when discrepancies between their expectation and the data were present.

Mee et al. (2013) continued the work of Clauser, Mee, et al. (2009) in examining the effect of manipulated data on the ratings of content experts. Mee et al. (2013) used the same items and data manipulation as the 2009 study in an effort to compare results; however, a procedural change was made in an attempt to examine whether the outcome of 2009 would have been different had different instructions been used. To this end, the authors told the judges about the 2009 study, informed them that some of the data had been manipulated, and explained that the judges should not utilize the feedback unless they believed it was accurate. The authors examined three standard setting panels and found that the modification of instructions caused the judges to make less use of the performance data than was the case in 2009. However, as in 2009, the extent of the changes made by judges was not substantially influenced by the accuracy of the feedback, leading to some additional concerns about the way in which judges are affected by feedback.

Clauser et al. (2013) conducted two studies in which they randomly assigned participants into two groups and provided each group with different type of feedback. The full data group received item-level examinee performance by decile and the percentage of examinees selecting each distractor option. The limited data group received only the distractor-level data. They found that the full data group had higher correlations between their final ratings and the conditional p-values of the items and a

decreased correlation between initial and final ratings. It is interesting to note that for both studies they found that correlations between the initial ratings across groups was stronger than the correlation between the initial ratings and conditional p-values, suggesting that the judges share some kind of view of item difficulty that cannot be fully explained by the empirically-derived item difficulties.

Finally, Margolis and Clauser (2014) investigated the impact of performance data on cut scores by examining the pre- and post-data recommendations of 18 independent standard setting panels. Following a round of rating items, judges were provided item-level examinee performance by decile and the percentage of examinees selecting each distractor option, and then provided their final ratings. The results indicated that the variability among judges decreased following the introduction of feedback and the post-data cut score recommendations were significantly different from the pre-data recommendations. Hurtz and Auerbach (2003) suggested that the introduction of feedback generally lowered cut score recommendations, but Margolis and Clauser (2014) found no support for such a claim.

Brandon (2004) conducted a lengthy review of the literature related to the modified-Angoff method and found several areas lacking. In particular, he cites the lack of research investigating the appropriate level of judge expertise and training methods. Brannon finds much of the research inconclusive and criticizes the low quality of some of the non-operational studies and the minimalist descriptions that often accompany standard setting research. He calls for richer descriptions of the methods used in

standard setting research, research comparing variations in the procedures, and experiments manipulating the steps of the method.

Chapter Three

Effects of Feedback based on the Difficulty of Two Forms

Introduction

Setting performance standards is a judgmental process involving human opinions and values as well as technical and empirical considerations. Although all cut score decisions are by nature arbitrary, they should not be capricious (AERA et al., 2009; Cizek, 2012; Shepard, 1979). Establishing a minimum passing standard is the technical expression of a policy decision. The information gained through standard setting studies informs these policy decisions. To this end, it is necessary to conduct robust examinations of standard setting studies in order to understand how the information gained from standard setting studies influences policy decisions.

Examining how information regarding item-level feedback influences the perceptions of item difficulty held by content experts is a subject that has not been studied extensively. However, the way in which information regarding item-level feedback influences content experts' decisions may hold extensive consequences with regard to setting an appropriate passing standard and the subsequent pass/fail or other categorical decisions. In particular, Hambleton et al. (2012) call for more research on the empirical results of performance data, specifically noting the interesting questions raised by providing the incorrect data in Clauser, Mee, et al. (2009). The current study seeks to examine how the item-level feedback provided to content experts affects the ratings they provide.

The primary research questions guiding this study are:

1. How does the item-level feedback provided to content experts influence the ratings they provide?
2. Does the difficulty of the standard setting form affect the ratings provided by content experts?

Participant Selection

Eligible participants (n=2,803) were sent an email requesting volunteers for a standard setting study. Eligible participants were all those who were certificate holders in good standing who passed the certification exam in 2012 or 2013 with a score of 600 or higher. Within a few days, 187 individuals had accepted the offer to participate and 177 completed training. A total of 171 judges provided ratings; 168 were fit for use following data cleaning.

Participant Training

All volunteers were required to complete a web-based training session of approximately 30-45 minutes in length. The group sessions were typically conducted over the course of a week and at varying times to account for volunteers in different time zones. Individual sessions were also available for those who were unable to participate in a group session. Additional assistance and technical support was available throughout the process by phone and email.

The primary focus of the training sessions was to familiarize judges with the modified-Angoff method and discuss a key concept of this method, that of a "minimally knowledgeable, yet certifiable candidate". The modified-Angoff method asks judges to determine the probability of a minimally-competent candidate answering a question correctly. Judges were asked to think of a physician they knew who they believe lacks the knowledge sufficient to be a board certified physician. They were then asked to think of a physician they knew who they believe would be considered barely qualified to be a board certified physician. In order to help conceptualize their understanding of a minimally qualified candidate, statements such as, "…this person would not be highly knowledgeable, but you would still be comfortable with them receiving the same certification that you have" were presented for consideration.

Judges were also provided an overview of the web-based rating software, including screenshots and instructions for accessing the website. Following a brief discussion on the mechanics of using the software, participants were provided an explanation of the concept of conditional p-values, or the percentage of minimally-qualified candidates who would answer a question correctly.

Finally, judges were shown a copy of a survey that would be administered once they completed rating each of the 120 items. Each survey question was reviewed to ensure that judges had a clear understanding of what was being asked and an understanding of how these questions factored into the standard setting process. This research was conducted during an operational standard setting exercise. The operational standard setting exercise utilized three distinct standard setting methods: (1) a modified-

Angoff method, (2) the Hofstee Method (1983), and (3) the Beuk Compromise Adjustment (1984). The responses to the survey questions are critical to the implementation of the Hofstee and Beuk methods, but are not within the scope of this research.

Data Collection

Item rating process

The asynchronous item rating process was designed to maximize participation by allowing judges to enter their ratings at their convenience during the rating window. This was accomplished through the use of a web-based software application that was available to the judges 24-hours a day during the rating window. For this study, the rating window was open for 18 days. The asynchronous nature of the process also eliminated the inconvenience and expense associated with requiring judges to travel. Although volunteers were not reimbursed for their time, they were recognized for their contributions with a framed acknowledgement.

The item rating process consisted of judges providing multiple data entries for 120 individual items. Judges begin by attempting to answer the question correctly and providing an initial difficulty rating for that question using a scale from 0-100. During training the judges were informed that this rating is the percentage of minimally-qualified candidates that they believe would get this question correct. Thus, 0 would be a difficult question and 100 would be an easy question. Once an answer and initial rating have been locked-in, judges are provided with the correct answer to the question and a conditional

p-value, which is a calculation of the percentage of examinees with ability levels at the current passing standard that would answer that question correctly. The conditional p-value is provided on the same 0-100 scale that judges use to rate the items. After receiving feedback regarding the correct answer and conditional p-value, judges are able to adjust their ratings for the item and submit a final rating. Judges are also asked a multiple choice question regarding their perception of the item and allowed to provide comments, but these issues are beyond the scope of this study.

Variables and Data Elements

The data returned contained 8 data elements, of which 3 are outside the scope of this research. The pertinent data elements are UserID, Form, Initial Rating, Final Rating, and Response Vector. The UserID variable is a unique identifier assigned to each judge that allows the responses collected from the standard setting software to be matched with the associated demographic information, which will be discussed shortly. The Form variable indicates the form of the standard setting questions. This study utilized two forms of items for judges; one form was considered easy and had a mean item difficulty scaled score of 200, while the other form was targeted to the exam's passing standard with a mean item difficulty scaled score of 390. The Initial Rating, Final Rating, and Response Vector variables exist for every item and are labeled sequentially according to the item sequence. For example, the Initial Rating variables in the dataset are labeled "InitialRating_1", "InitialRating_2", " InitialRating_3"...etc. Therefore, there exist initial ratings, final ratings, and response vectors for each item rated by each judge. The Initial Rating variable is the 0-100 rating provided by each judge on each item before they

received feedback. The Final Rating variable is the 0-100 rating provided by each judge on each item after they received feedback. The Response Vector variable is a 0-1 scoring for every item based on whether the judge answered the question correctly (1) or incorrectly (0) before they received any feedback.

As previously mentioned, the UserID variable allows for matching rating sets to the appropriate demographic information. The demographic information for each judge included gender, medical degree (i.e. MD or DO), score on the last exam, and whether they were a candidate for initial certification or recertification on their last exam. This demographic information is used to ensure that there is adequate representation and that judge selection is not biased.

Data Cleaning

As is typically the case, the results from the item rating process were returned with missing data points as well as instances of misuse of the rating scale by judges. The most common issue was that of misuse of the rating scale. Judges often provided single-digit ratings of item difficulty. On the 0-100 scale, a single-digit response would represent a question so difficult that less than 10 percent of minimally qualified examinees would answer it correctly. Judges often included comments indicating that they had made some kind of mistake in providing the rating for specific items. In other cases it was relatively obvious that a typographical error existed. In each of these cases, the single-digit ratings were transformed onto the 100-point scale by multiplying the rating provided by 10. Although there is no rule governing the extent to which these transformations were tolerated, if a judge had multiple instances (typically more than 10)

they were removed from the dataset due to significant misuse of the scale. It was also common for judges to provide ratings for only a portion of the items. If any number of the ratings were missing that judge was removed for having an incomplete dataset.

For this study, five judges provided a single initial and final rating using a 10-point scale, while another judge provided 6 initial and 6 final ratings using a 10-point scale. In addition, seven judges provided a single initial rating using a 10-point scale, one judge provided two initial ratings using a 10-point scale, and a third provided three initial ratings using a 10-point scale. These judges used a 10-point scale for their initial ratings, but corrected themselves and used the proper 100-point scale for their final ratings. Conversely, one judge provided a single final rating using a 10-point scale and another provided two final ratings using a 10-point scale. In each of these instances, the single-digit scores were transformed onto the 100-point scale by multiplying the rating provided by 10.

There were four judges who were completely removed from this dataset. Two judges provided all 120 items on a 10-point scale and were removed for incorrect use of the scale. One judge provided only five ratings, while another provided 68. Both of these judges were removed for providing incomplete datasets.

Creating conditional p-values

The modified-Angoff method utilized here is a content-based method for recommending a passing standard that asks content experts to examine each item and determine the probability of a "minimally competent examinee" answering a question

correctly. Judges are commonly provided some form of past examinee performance data to assist in their decision-making process. Each of the examinations under consideration here was scored using the dichotomous Rasch model (Rasch, 1960). The Rasch model is a logistic model of latent traits that provides person measures and item difficulties in log-odds units, commonly referred to as logits. There are two options available for providing performance data to judges regarding the probability of a minimally competent candidate answering a question correctly: (1) aggregate the item-level responses only from previous examinees who scored at the passing standard, and (2) calculate conditional p-values for each item. Using the responses from previous examinees may provide a more accurate reflection of the ability of minimally competent examinees, but there is typically a dearth of examinees who scored at the passing standard for each item under examination. Therefore, the typical action is to create conditional p-values for each item. Conditional p-values are calculations of the percentage of candidates with ability estimates at the passing standard expected to get the question correct.

Providing judges a calculation of conditional p-values rather than overall item difficulty is done in order to give standard setting judges a more accurate view of how minimally competent examinees would actually perform on this item rather than relying on judges trying to estimate a probability of success for each item based on their own sense of how a minimally competent examinee might perform. However, the overall item difficulty is used to calculate the conditional p-values. The transformation of the overall item difficulty into a conditional p-value was accomplished using the following formula:

$$\Pr\{\ \beta_{MPS} = 1\} = \frac{e^{MPS-\delta i}}{1 + e^{MPS-\delta i}}$$

Where:

$\Pr\{\ \beta_{MPS} = 1\}$	is the probability of a correct response by a minimally qualified candidate
MPS	is the calibration of the minimum passing standard
δi	is the difficulty of item i
e	is the base of the natural logarithm

The overall item difficulty calibration in logits from previous exam administrations is subtracted from the calibration in logits for the minimum passing score to produce a conditional difficulty measure of each item for a minimally qualified candidate. The base of the natural logarithm for this new calibration is divided by 1 + the base of the natural logarithm, which produces a conditional probability. This conditional probability is multiplied by 100 in order to return a percentage in a whole number that judges can readily understand. This percentage is referred to as the conditional p-value of an item.

Methods

The primary research question for this study is, "How does item-level feedback provided to content experts influence the ratings they provide?" In order to fully explore this question, I employ several strategies designed to examine the various ways in which the final ratings of judges are influenced by the feedback provided.

There were two types of feedback provided to judges in this standard setting exercise: (1) conditional p-value of item difficulty, and (2) whether the judge was able to correctly answer the question. Furthermore, this standard setting study utilized two forms

of items for judges. One form was considered easy and had a mean item difficulty scaled score of 200, while the other form was targeted to the exam's passing standard with a mean item difficulty scaled score of 390. Therefore, in addition to the primary research question investigating item-level feedback, a second research question asks, "does the difficulty of the form of the exam affect the ratings provided by judges?"

In order to examine these two research questions, for the entire cohort of judges, as well as for each form of the exam, I will perform a one-way repeated measures ANOVA to determine whether the means for the initial and final ratings are significantly different. Next, in order to examine the effects of the conditional p-value feedback, I will calculate a Pearson product-moment correlation coefficient for each judge's initial ratings with the conditional p-values and their final ratings with the conditional p-values. A stronger correlation coefficient would suggest that the judge adjusted their rating to be more in line with the conditional p-values. Finally, in order to examine the effect of whether the judge answered the question correctly, I will examine their average change in ratings for questions answered correctly and as well as the average change for those questions answered incorrectly. A paired-samples t-test will be employed to determine whether the mean change for correct answers is significantly different from the mean change for incorrect answers.

It is also important to examine whether there are differences in the ratings provided based on demographic variables. Therefore, for each form of the exam, repeated measures factorial (mixed) ANOVA tests will be conducted to determine whether there are any interaction effects based on certain demographic variables: gender,

medical degree (MD, DO), and certification status (initial certifiers and candidates for recertification).

Results

The results of the one-way ANOVA show that the judges' ratings were significantly affected by the feedback provided, $F(1, 167) = 112.7$, $p < .001$. With regard to the effects of the conditional p-value feedback, the correlation of the judges' ratings with the conditional p-values provided increased following the introduction of the feedback (Table 3.1). Of the 168 judges, 14 (8.3%) did not change their ratings to a degree that it altered their correlation coefficient, 100 (59.5%) changed their ratings such that their correlation coefficient increased by less than .1, 44 (26.2%) changed their ratings such that their correlation coefficient increased by .1 or more, and 10 (6.0%) changed their ratings such that their correlation coefficient decreased.

Table 3.1.
Summary of Correlation with Conditional P-value feedback

	Initial Rating		Final Rating	
	Mean	SD	Mean	SD
Both Forms	0.60	0.21	0.67	0.21
Easy Form	0.59	0.24	0.65	0.25
Hard Form	0.62	0.16	0.70	0.17

Although the final ratings were more strongly correlated with the conditional p-value feedback provided, the initial ratings remained strongly associated with their final ratings as seen in figures 3.1, 3.2, and 3.3. These figures show each judge's initial rating

(before feedback is provided) on the X-axis plotted against their final rating (following feedback) on the Y-axis for each item. A linear regression line and associated R-square value are also provided.

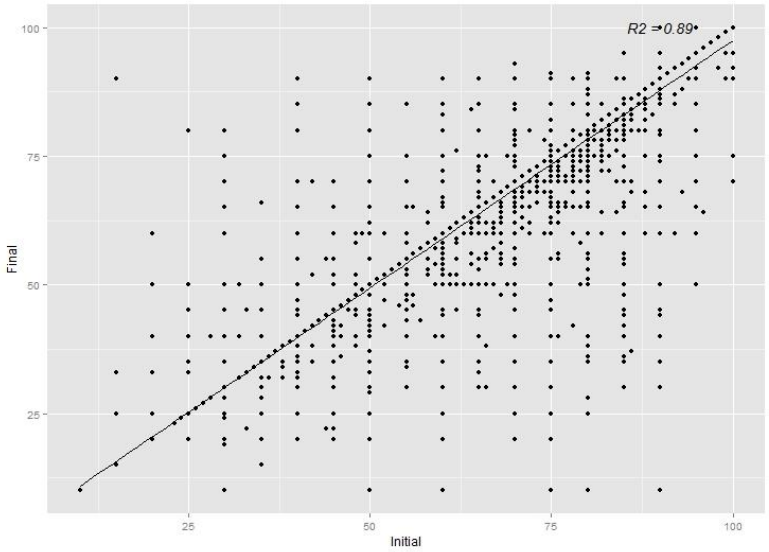

Figure 3.1. Plot of Initial Rating with Final Rating.

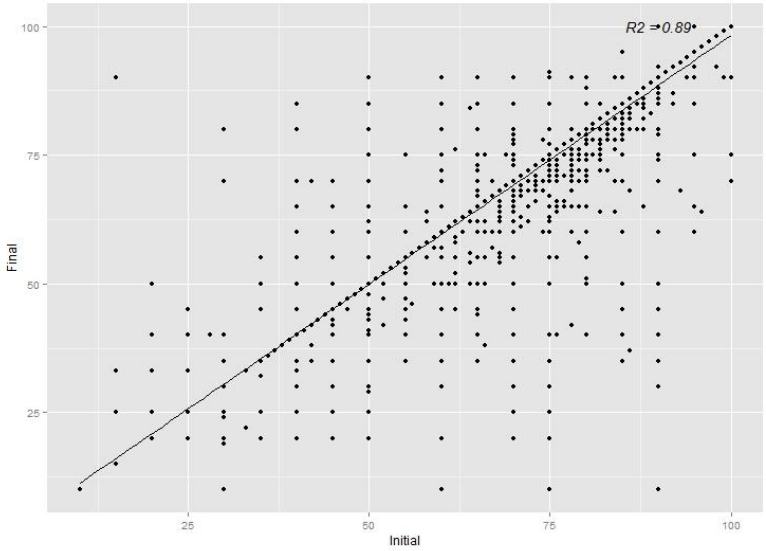

Figure 3.2. Plot of Initial Rating with Final Rating (EASY FORM)

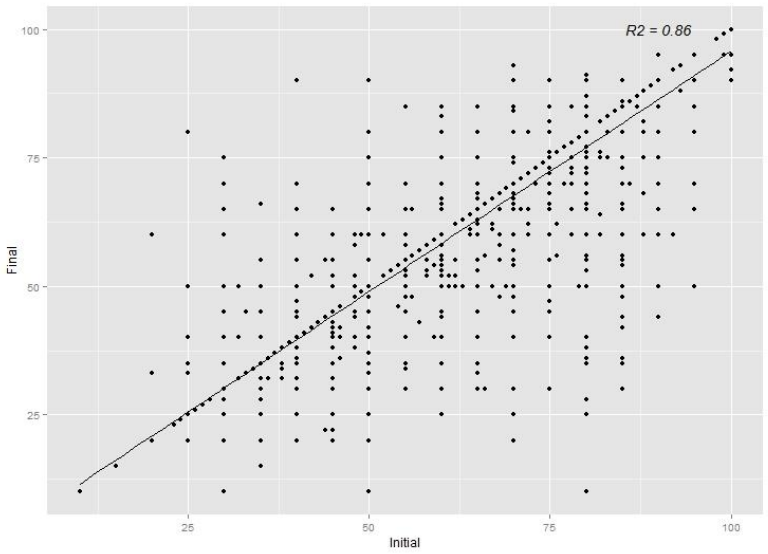

Figure 3.3. Plot of Initial Rating with Final Rating (HARD FORM)

With regard to the effect of judges answering the question correctly, on average those receiving the Easy form changed their rating to a significantly greater degree when answering a question incorrectly (*M*=4.6, *SE* =.42) as opposed to answering a question correctly (*M*=.96, *SE* =.12), *t* (84) = -8.92, *p* <.001. Similarly, on average those receiving the Hard form changed their rating to a significantly greater degree when answering a question incorrectly (*M*=5.3, *SE* =.47) as opposed to answering a question correctly (*M*=1.8, *SE*=.19), *t* (82) = -7.63, *p* <.001.

An examination of the interaction effects of demographic variables showed there was no significant interaction effect of gender, indicating that the ratings provided by male and female judges were generally the same on both the Easy form, *F*(1, 83) = 1.9, *p*=.17 and the Hard form, *F*(1, 81) = .05, *p*=.82. Similarly, there was no significant interaction effect of medical degree (MD or DO), indicating that the ratings provided by those with allopathic medical training and those with osteopathic medical training were

generally the same on both the Easy form, $F(1, 83) = 3.1$, $p=.08$ and the Hard form, $F(1, 81) = .66$, $p=.42$. Finally, there was no significant interaction effect of certification status, indicating that the ratings provided by those who has just completed their initial certification and those who had recertified with at least 7 year of prior practice were generally the same on both the Easy form, $F(1, 83) = .04$, $p=.84$ and the Hard form, $F(1, 81) = .19$, $p=.66$.

Discussion

This study sought to explore how item-level feedback provided to content experts affected the ratings they provide and whether the difficulty of the standard setting form affected those same ratings. The results indicate that judges did indeed utilize the conditional p-value feedback; however, although these results are statistically significant they do not seem to be practically significant. Figures 3.1, 3.2, and 3.3 show that the association between the initial ratings and final ratings remain strong even after feedback, suggesting that judges tend to primarily rely on their innate sense of item difficulty rather than the conditional p-values provided. Further to this point, 10 judges altered their ratings in such a way that their correlation coefficient decreased; meaning that after having been provided feedback, they made a conscious decision to adjust their ratings in the opposite direction of the feedback.

It would seem that the more important feedback mechanism was whether or not the judges were able to correctly answer the question. If a judge answered a question incorrectly they were more likely to change their rating to be closer in line with the conditional p-value provided; conversely, if they answered the question correctly they

were unlikely to make much of a change at all. These results hold for those receiving the Easy form as well as those receiving the Hard form. Furthermore, those judges who received the Hard form made larger changes in their ratings than those who received the Easy form. This suggests that the difficulty of the standard setting form used does affect the ratings provided, particularly with regard to whether the items are appropriately targeted to the ability level of the judges. As previously mentioned, the Easy form was constructed to have an item mean targeting a scaled score of 200 and the Hard form was constructed to have an item mean targeting the exam's current passing standard of 390. The Hard form was much more representative of an appropriate standard setting form, but in both cases if the judges were not able to answer the questions then they relied more heavily on the conditional p-value feedback. Therefore, when constructing standard setting forms care must be taken to ensure that both the standard setting form and the ability of the judges are sufficiently matched in order to provide informative ratings.

One of the primary criticisms of the Angoff method is that judges are unable to accurately estimate the difficulty of items for minimally qualified candidates (Busch & Jaeger, 1990; Clauser et al., 2002; Cross et al., 1984; Impara & Plake, 1998; Reckase, 2000). This study finds that although the correlation of judge ratings to conditional p-values before the introduction of feedback was not high, the introduction of feedback did not increase the correlation to a practically significant degree; the judges seemed relatively confident in their initial ratings. However, I would argue that this finding does not support the view of Shepard et al. (1993) in determining that the Angoff method is fundamentally flawed. Rather, I contend that the issue is more that of judge selection

criteria and ensuring that those participating in standard setting exercise be appropriately qualified and able to correctly answer the questions.

An additional criticism is that the judges are too reliant on the feedback provided (Hurtz & Auerbach, 2003; Maurer & Alexander, 1992; Truxillo et al., 1996). The results here indicate that this is clearly not the case. Judges are typically instructed to incorporate the feedback as a supplement to their opinion as a content expert. It seems that this is exactly what they're doing. However, this study had the luxury of a large sample size and lack of group effect. In a group setting the feedback may serve as a convenient point upon which the judges may converge, but that is an issue of group effect more than being overly reliant on the feedback.

Previous studies have found that inter-rater agreement increased between rounds following rater discussion, but this discussion did not increase the correlation between ratings and conditional p-values. The correlation between ratings and conditional p-values did not increase until the introduction of some form of empirical item-level feedback. The ability of a group of judges to come to a common consensus regarding item difficulty is often seen as one of the benefits of the Angoff method; however, it is also possible that a strong personality in a group could sway the ratings. Clauser et al. (2002) found a substantial group effect and noted that discussion without feedback improved judge agreement within groups, but not between groups. The inability of groups of judges to provide consistent results across groups is one of the primary criticisms of the Angoff method and led to Clauser, Margolis, and Clauser (2014) and Hambleton et al. (2012) recommending that standard setting panels be conducted with

multiple groups. This study was conducted asynchronously, eliminating the confounding inter-rater effect and allowing for an analysis of the perceptions of each individual judge.

Another notable difference in this study is that the feedback followed each item rather than being provided between rounds. Typically, judges rate all items, hold a discussion, examine feedback, and then provide a final rating. The methodology utilized here whereby judges provided an initial rating, received feedback, and then provided a final rating on an item before moving onto the next item allows for an analysis of each item independent of the other items. It may also allow judges to calibrate their internal sense of item difficulty with the conditional p-values early in the exercise rather than having to recalibrate between rounds.

Some researchers (Cizek, 1996; Hambleton & Pitoniak, 2006; Kane, 1994; Loomis, 2012; Raymond & Reid, 2001) have proposed that all relevant stakeholders for an examination should be invited to participate as judges in standard setting exercises. This study demonstrates that inviting individuals who are not content experts, and likely do not have the ability to correctly answer the questions, would negatively affect the results of a standard setting exercise and the resulting recommended cut score. For example, a member of the general public invited to participate in a standard setting study for medical licensure would lean so heavily on the feedback provided that it would be of little sense to have them involved. In this same scenario, if the feedback provided was a conditional p-value based on the ability of a minimally qualified candidate, recalling that the calculation to determine the ability of a minimally qualified candidate is based on the

current passing standard, the recommended passing score for this judge would be a self-fulfilling prophecy of retaining the current standard.

To my knowledge, there are no published research findings on standard setting exercises that utilize an asynchronous design. Harvey and Way (1999) and Harvey (2000) discuss the creation of a web-based application to conduct standard setting exercises and the differences in how judges felt about their experience using the web-based application compared to an in-person session, while MacCann and Stanley (2010) outline some of the potential benefits of a web-based standard setting exercise. The design utilized here allows for an examination of the effects of feedback on individual judges without the influence of a group effect. Future standard setting studies, both operational and research-based, should consider utilizing a similar design. The ability to eliminate the group effect and isolate the ratings and subsequent recommended cut score of an individual judge, while at the same time significantly increasing the number of judges involved, should lend itself to an increase in the reliability and validity of standard setting exercises. However, there clearly needs to be additional research conducted in order to support this claim.

The results of this study contribute to the body of evidence on the effects of feedback on ratings, the effects of form difficulty, and the criteria for judge selection. The outcome of high-stakes testing determines whether an individual receives a diploma, gets into certain colleges and universities, and is granted entry into certain professions. In medical licensure, the outcome can literally be a matter of life or death if an unqualified physician is granted license to perform certain procedures. Cizek (2012)

notes that the determination of cut scores also influences decisions about whether or not death penalty sentences should be carried out. With such broad and sweeping consequences, it is incumbent upon those conducting standard setting exercises to utilize the most rigorous methods available. In his 2004 review of the literature, Paul Brandon concludes that those conducting standard setting research are "not attending to the most rudimentary prescriptions about describing methods in sufficient detail to evaluate or replicate standard setting studies" (p.80). Furthermore, Brandon rues the "lack of a comprehensive program of standard setting research" (p.80). Clearly, research in this field continues to be necessary if we are ever to come to a consensus on the appropriate methods for setting passing standards.

Conclusions

Limitations

The primary limitation of this study is that it is correlational research and, while useful to help uncover the relationship between variables, does not provide and conclusive evidence for causation and often leads to more questions than answers. There is also an issue of whether the results of this study are generalizable since the judges were all board-certified physicians and the item-level feedback may affect these highly-trained content experts differently than content experts in other fields. Additionally, there is an issue that there were no common items between the Easy and Hard forms of the exam. The ability to examine a significant number of common items between two forms may have provided a more nuanced look at how the difficulty of the form influences the ratings provided by content experts.

A final limitation is that this standard setting exercise was conducted asynchronously and not in-person as is customary in Angoff-style standard setting. However, I see this as a benefit rather than a limitation. The ability of a group of judges to come to a common consensus regarding item difficulty is often seen as one of the benefits of the Angoff method. For this study there is no inter-judge agreement and the ability of one strong personality to sway the opinion of the group is not an issue. The asynchronous nature of this study also allows for the analysis of each individual judge without the additional confounding variable of judge inter-judge agreement. There is a fundamental difference in the process using an asynchronous method from the traditional in-person method. The in-person method focuses on discussion and consensus building, while the asynchronous method focuses on each judge constructing their own individual idea of a minimally qualified candidate. Although the outcome of the process is the same, recommending a minimum passing standard, this difference is critical to the analysis of the feedback mechanism and investigating how judges come to understand the idea of a minimally qualified candidate.

Educational Significance

The modified-Angoff method remains one of the most popular methods for setting performance standards in testing and assessment (Hurtz & Auerbach, 2003; Plake & Cizek, 2012). It is common practice to provide content experts with feedback regarding the item-level difficulties; however, it is unclear how this feedback affects the ratings and recommendations of content experts. Recent research seems to indicate mixed results, noting that the feedback given to judges may or may not alter their ratings

depending on the type of data provided, when the data was provided, and how judges collaborated within groups and between groups. The research proposed here seeks to examine issues related to the effects of item-level feedback on the ratings provided by judges. The results of this research may hold implications for how standard setting studies are conducted with regard to the difficulty and ordering of items, the ability level of content experts invited to participate in these studies, and the type of feedback that is provided to judges.

Standard setting methods that utilize judges and ask them to conceptualize a minimally qualified candidate (e.g. Angoff, Bookmark, etc.) will always have an issue regarding variance among what judges consider to be minimally qualified. This will be true in the realm of medical testing, K-12 testing, or any other field that utilized standardized tests. In high-stakes testing, setting performance standards is of critical importance and it is imperative that the utmost care be taken to ensure that standard setting exercises are conducted with the strongest theoretical and empirical foundation possible. I hope this research will add to the body of literature regarding the way in which standard setting studies are conducted and that researchers will begin to form a consensus around best-practices in setting performance standards.

Chapter Four

Effects of Feedback based on the Ability of the Judge

Introduction

Setting performance standards is a judgmental process involving human opinions and values as well as technical and empirical considerations. Although all cut score decisions are by nature arbitrary, they should not be capricious (AERA et al., 2009; Cizek, 2012; Shepard, 1979). Establishing a minimum passing standard is the technical expression of a policy decision. The information gained through standard setting studies informs these policy decisions. To this end, it is necessary to conduct robust examinations of standard setting studies in order to understand how the information gained from standard setting studies influences policy decisions.

Examining how information regarding how item-level feedback influences the perceptions of item difficulty held by content experts is a subject that has not been studied extensively. However, the way in which information regarding item-level feedback influences content experts' decisions may hold extensive consequences with regard to setting an appropriate passing standard and the subsequent pass/fail or other categorical decisions. In particular, Hambleton et al. (2012) call for more research on the empirical results of performance data, specifically noting the interesting questions raised by providing the incorrect data in Clauser, Mee, et al. (2009). The current study seeks to examine how the item-level feedback provided to content experts affects the ratings they provide.

The primary research questions guiding this study are:

1. How does the item-level feedback provided to content experts influence the ratings they provide?
2. Does the ability level of content experts affect the ratings they provide?

Participant Selection

Eligible participants (n=2,395) were sent an email requesting volunteers for the standard setting study. Eligible participants were all those who were certificate holders in good standing who passed the most recent certification exam with a score of 600 or higher (n=1,838) or who passed with a score of 390 or 400 (n=557). Within a few days, 287 individuals had accepted the offer and 144 were subsequently selected to participate in the standard setting study. Of these 144 volunteers, 122 completed the training and 91 provided ratings. Following data cleaning, 81 were fit for use.

Participant Training

All volunteers were required to complete a web-based training session of approximately 30-45 minutes in length. The group sessions were typically conducted over the course of a week and at varying times to account for volunteers in different time zones. Individual sessions were also available for those who were unable to participate in a group session. Additional assistance and technical support was available throughout the process by phone and email.

The primary focus of the training sessions was to familiarize judges with the modified-Angoff method and discuss a key concept of this method, that of a "minimally knowledgeable, yet certifiable candidate". The modified-Angoff method asks judges to determine the probability of a minimally-competent candidate answering a question correctly. Judges were asked to think of a physician they knew who they believe lacks the knowledge sufficient to be a board certified physician. They were then asked to think of a physician they knew who they believe would be considered barely qualified to be a board certified physician. In order to help conceptualize their understanding of a minimally qualified candidate, statements such as, "…this person would not be highly knowledgeable, but you would still be comfortable with them receiving the same certification that you have" were presented for consideration.

Judges were also provided an overview of the web-based rating software, including screenshots and instructions for accessing the website. Following a brief discussion on the mechanics of using the software, participants were provided an explanation of the concept of conditional p-values, or the percentage of minimally-qualified candidates who would answer a question correctly.

Finally, judges were shown a copy of a survey that would be administered once they completed rating each of the 120 items. Each survey question was reviewed to ensure that judges had a clear understanding of what was being asked and an understanding of how these questions factored into the standard setting process. This research was conducted during an operational standard setting exercise. The operational standard setting exercise utilized three distinct standard setting methods: (1) a modified-

Angoff method, (2) the Hofstee Method (1983), and (3) the Beuk Compromise Adjustment (1984). The responses to the survey questions are critical to the implementation of the Hofstee and Beuk methods, but are not within the scope of this research.

Data Collection

Item rating process

The asynchronous item rating process was designed to maximize participation by allowing judges to enter their ratings at their convenience during the rating window. This was accomplished through the use of a web-based software application that was available to the judges 24-hours a day during the rating window. The window for this study was open for 19 days. The asynchronous nature of the process also eliminated the inconvenience and expense associated with requiring judges to travel. Although volunteers were not reimbursed for their time, they were recognized for their contributions with a framed acknowledgement.

The item rating process consisted of judges providing multiple data entries for 120 individual items. Judges begin by attempting to answer the question correctly and providing an initial difficulty rating for that question using a scale from 0-100. During training the judges were informed that this rating is the percentage of minimally qualified candidates that they believe would get this question correct. Thus, 0 would be a difficult question and 100 would be an easy question. Once an answer and initial rating have been locked-in, judges are provided with the correct answer to the question and a conditional

p-value, which is a calculation of the percentage of examinees with ability levels at the current passing standard that would answer that question correctly. The conditional p-value is provided on the same 0-100 scale that judges use to rate the items. After receiving feedback regarding the correct answer and conditional p-value, judges are able to adjust their ratings for the item and submit a final rating. Judges are also asked a multiple choice question regarding their perception of the item and allowed to provide comments, but these issues are beyond the scope of this study.

Variables and Data Elements

The data returned contained 8 data elements, of which 3 are outside the scope of this research. The pertinent data elements are UserID, Form, Initial Rating, Final Rating, and Response Vector. The UserID variable is a unique identifier assigned to each judge that allows the responses collected from the standard setting software to be matched with the associated demographic information, which will be discussed shortly. The Form variable indicates which version of the standard setting items the rater saw. For this study all judges saw the same form. The Initial Rating, Final Rating, and Response Vector variables exist for every item and are labeled sequentially according to the item sequence. For example, the Initial Rating variables in the dataset are labeled "InitialRating_1", "InitialRating_2", "InitialRating_3"...etc. Therefore, there exist initial ratings, final ratings, and response vectors for each item rated by each judge. The Initial Rating variable is the 0-100 rating provided by each judge on each item before they received feedback. The Final Rating variable is the 0-100 rating provided by each judge on each item after they received feedback. The Response Vector variable is a 0-1 scoring

for every item based on whether the judge answered the question correctly (1) or incorrectly (0) before they received any feedback.

As previously mentioned, the UserID variable allows for matching rating sets to the appropriate demographic information. The demographic information for each judge included gender, medical degree (i.e. MD or DO), score on the last exam, and whether the judge was a candidate for initial certification or recertification on the last exam. This demographic information is used to ensure that there is adequate representation and that judge selection is not biased.

Data Cleaning

As is typically the case, the results from the item rating process were returned with missing data points as well as instances of misuse of the rating scale by judges. The most common issue was that of misuse of the rating scale. Judges often provided single-digit ratings of item difficulty. On the 0-100 scale, a single-digit response would represent a question so difficult that less than 10 percent of minimally qualified examinees would answer it correctly. Judges often included comments indicating that they had made some kind of mistake in providing the rating for specific items. In other cases it was relatively obvious that a typographical error existed. In each of these cases, the single-digit ratings were transformed onto the 100-point scale by multiplying the rating provided by 10. Although there is no rule governing the extent to which these transformations were tolerated, if a judge had multiple instances (typically more than 10) they were removed from the dataset due to significant misuse of the scale. It was also

common for judges to provide ratings for only a portion of the items. If any number of the ratings were missing that judge was removed for having an incomplete dataset.

For this study, one judge provided a single initial and final rating using a 10-point scale. Another provided two initial ratings and two final ratings using a 10-point scale. A third judge provided five initial and five final ratings using a 10-point scale. Four judges provided a single initial rating using a 10-point scale, but corrected themselves and used the proper 100-point scale for their final ratings. In each of these instances, the single-digit scores were transformed onto the 100-point scale by multiplying the rating provided by 10.

There were 22 total judges removed from this dataset. Eleven judges were removed for not providing enough ratings. They provided 2, 4, 6, 6, 15, 19, 32, 49, 58, 61, and 70 ratings, respectively. An additional eleven judges were removed from the dataset for improper use of the rating scale. Four of these judges provided a significant number of ratings using a 10-point scale; one judge provided 10 ratings, another 20, a third 47, and a fourth 64. Two judges were removed for providing all 120 ratings on a 10-point scale and four were removed for providing all 120 ratings using a 5-point scale. The final judge was removed for improper use of the scale provided a rating of "50" for 119 of the 120 items.

Creating conditional p-values

The modified -Angoff method utilized here is a content-based method for recommending a passing standard that asks content experts to examine each item and

determine the probability of a "minimally competent examinee" answering a question correctly. Judges are commonly provided some form of past examinee performance data to assist in their decision-making process. Each of the examinations under consideration here was scored using the dichotomous Rasch model (Rasch, 1960). The Rasch model is a logistic model of latent traits that provides person measures and item difficulties in log-odds units, commonly referred to as logits. There are two options available for providing performance data to judges regarding the probability of a minimally competent candidate answering a question correctly: (1) aggregate the item-level responses only from previous examinees who scored at the passing standard, and (2) calculate conditional p-values for each item. Using the responses from previous examinees may provide a more accurate reflection of the ability of minimally competent examinees, but there is typically a dearth of examinees who scored at the passing standard for each item under examination. Therefore, the typical action is to create conditional p-values for each item. Conditional p-values are calculations of the percentage of candidates with ability estimates at the passing standard expected to get the question correct.

Providing judges a calculation of conditional p-values rather than overall item difficulty is done in order to give standard setting judges a more accurate view of how minimally competent examinees would actually perform on this item rather than relying on judges trying to estimate a probability of success for each item based on their own sense of how a minimally competent examinee might perform. However, the overall item difficulty is used to calculate the conditional p-values. The transformation of the overall item difficulty into a conditional p-value was accomplished using the following formula:

$$\Pr\{\ \beta_{MPS} = 1\} = \frac{e^{MPS-\delta i}}{1 + e^{MPS-\delta i}}$$

Where:

$\Pr\{\ \beta_{MPS} = 1\}$	is the probability of a correct response by a minimally qualified candidate
MPS	is the calibration of the minimum passing standard
δi	is the difficulty of item i
e	is the base of the natural logarithm

The overall item difficulty calibration in logits from previous exam administrations is subtracted from the calibration in logits for the minimum passing score to produce a conditional difficulty measure of each item for a minimally qualified candidate. The base of the natural logarithm for this new calibration is divided by 1 + the base of the natural logarithm, which produces a conditional probability. This conditional probability is multiplied by 100 in order to return a percentage in a whole number that judges can readily understand. This percentage is referred to as the conditional p-value of an item.

Methods

The primary research question for this study is, "How does item-level feedback provided to content experts influence the ratings they provide?" In order to fully explore this question, I employ several strategies designed to examine the various ways in which the final ratings of judges are influenced by the feedback provided.

There were two types of feedback provided to judges in this standard setting study: (1) conditional p-value of item difficulty, and (2) whether the judge was able to correctly answer the question. Furthermore, this standard setting study included two

cohorts of subject matter experts (SMEs): Highly Qualified SMEs who scored 600 or above on the last examination and Minimally Qualified SMEs who just met the passing standard with a score of 390 or 400. Therefore, in addition to the primary research question investigating item-level feedback, a second research question asks, "does the ability level of the judges affect the ratings they provide?"

In order to examine these two research questions I will perform a one-way repeated measures ANOVA to determine whether the means for the initial and final ratings are significantly different. Next, in order to examine the effects of the conditional p-value feedback, I will calculate a Pearson product-moment correlation coefficient for each judge's initial ratings with the conditional p-values and their final ratings with the conditional p-values. A stronger correlation coefficient would suggest that the judge adjusted their rating to be more in line with the conditional p-values. Finally, in order to examine the effect of whether the judge answered the question correctly, I will examine their average change in ratings for questions answered correctly and as well as the average change for those questions answered incorrectly. A paired-samples t-test will be employed to determine whether the mean change for correct answers is significantly different from the mean change for incorrect answers.

It is also important to examine whether there are differences in the ratings provided based on demographic variables. Therefore, repeated measures factorial (mixed) ANOVA tests will be conducted to determine whether there are any interaction effects based on certain demographic variables: gender, medical degree (MD, DO),

certification status (initial certifiers and candidates for recertification), and Subject Matter Expert group (390/400, 600+).

Results

The results of the one-way ANOVA show that the judges' ratings were significantly affected by the feedback provided, $F(1, 80) = 31.1$, $p < .001$. With regard to the effects of the conditional p-value feedback, the correlation of the judges' ratings with the conditional p-values provided increased following the introduction of the feedback (Table 4.1). Of the 81 judges, 6 (7.4%) did not change their ratings to a degree that it altered their correlation coefficient, 21 (25.9%) changed their ratings such that their correlation coefficient increased by less than .1, 53 (65.4%) changed their ratings such that their correlation coefficient increased by .1 or more, and 1 (1.2%) changed their ratings such that their correlation coefficient decreased.

Table 4.1.
Summary of Correlation with Conditional P-value feedback

	Initial Rating		Final Rating	
	Mean	SD	Mean	SD
Both Groups	0.37	0.16	0.54	0.19
High SMEs	0.38	0.15	0.54	0.18
Minimal SMEs	0.33	0.21	0.54	0.24

Although the final ratings were more strongly correlated with the conditional p-value feedback provided, the initial ratings remained strongly associated with their final ratings as seen in figures 4.1, 4.2, and 4.3. These figures show each judge's initial rating

(before feedback is provided) on the X-axis plotted against their final rating (following feedback) on the Y-axis for each item. A linear regression line and associated R-square value are also provided.

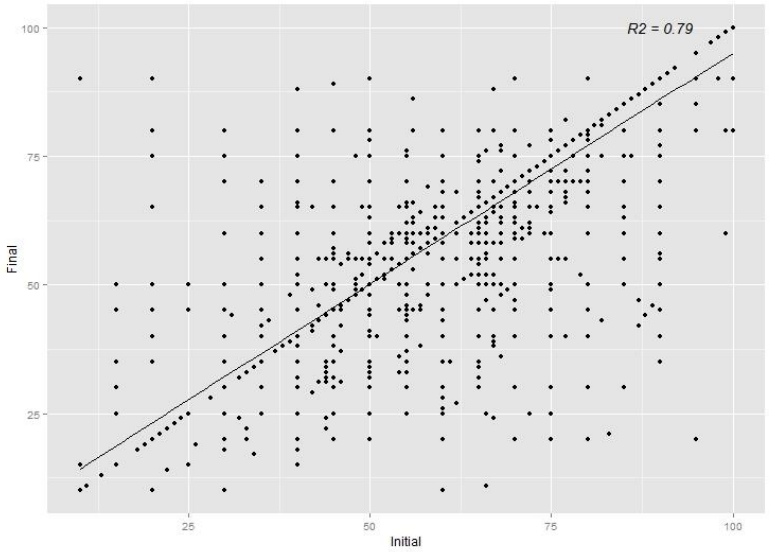

Figure 4.1. Plot of Initial Rating with Final Rating.

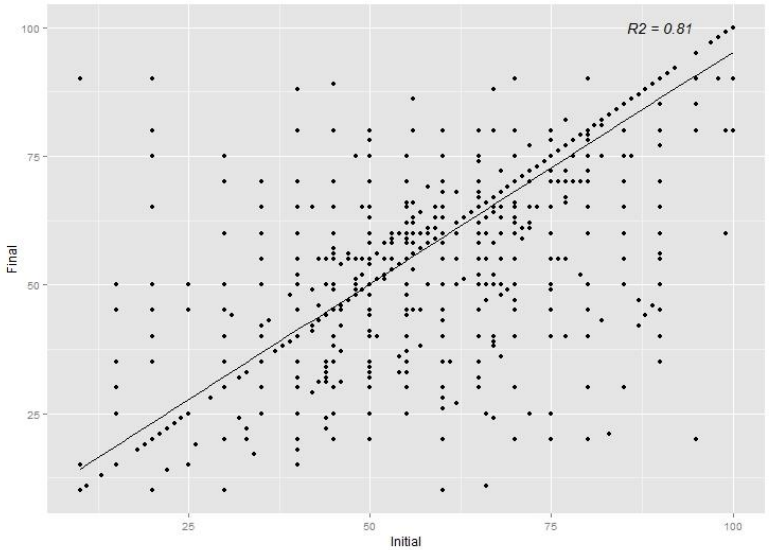

Figure 4.2. Plot of Initial Rating with Final Rating (High SMEs)

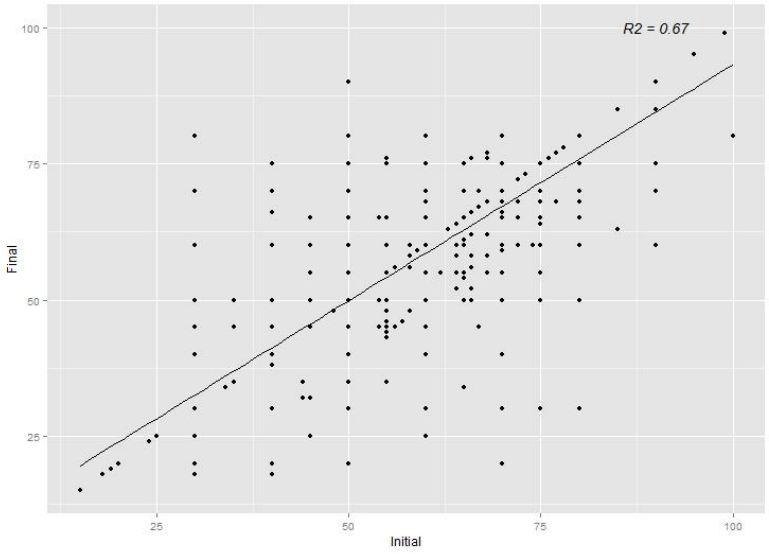

Figure 4.3. Plot of Initial Rating with Final Rating (Minimal SMEs)

With regard to the effect of judges answering the question correctly, on average the Highly Qualified SMEs changed their rating to a significantly greater degree when answering a question incorrectly (*M*=4.9, *SE* =.44) as opposed to answering a question correctly (*M*=2.2, *SE*=.22), *t* (65) = -6.57, *p* <.001. Similarly, on average the Minimally Qualified SMEs changed their rating to a significantly greater degree when answering a question incorrectly (*M*=6.0, *SE*=1.2) as opposed to answering a question correctly (*M*=2.4, *SE*=.62), *t* (14) = -3.25, *p* <.01.

An examination of the interaction effects of demographic variables showed there was no significant interaction effect of gender, indicating that the ratings provided by male and female judges were generally the same, *F*(1, 79) = .13, *p*=.72. Similarly, there was no significant interaction effect of medical degree (MD or DO), indicating that the ratings provided by those with allopathic medical training and those with osteopathic medical training were generally the same, *F*(1, 79) = 1.14, *p*=.29. In addition, there was

no significant interaction effect of certification status, indicating that the ratings provided by those who has just completed their initial certification and those who had recertified with at least 7 year of prior practice were generally the same, $F(1, 79)= .26$, $p=.61$. Finally, there was no significant interaction effect of subject matter expert group, indicating that the ratings provided by Highly Qualified SMEs and Minimally Qualified SMEs were generally the same, $F(1, 79)= .57$, $p=.45$.

Discussion

This study sought to explore how item-level feedback provided to content experts affected the ratings they provide and whether the ability level of subject matter experts affected those same ratings. The results indicate that judges did indeed utilize the conditional p-value feedback; however, although these results are statistically significant they do not seem to be practically significant. Figures 4.1, 4.2, and 4.3 show that the association between the initial ratings and final ratings remain strong even after feedback, suggesting that judges tend to primarily rely on their innate sense of item difficulty rather than the conditional p-values provided. However, these figures also illustrate that Highly Qualified SMEs exhibited much less variance between their initial and final ratings than the Minimally Qualified SMEs. Table 4.1 shows a smaller standard deviation for the Highly Qualified SMEs, indicating greater agreement within that group. Taken together, these results suggest that Highly Qualified SMEs have a slightly more accurate initial sense of item difficulty, although both groups provided similar ratings following the introduction of the conditional p-value feedback.

It would seem that the more important feedback mechanism was whether or not the judges were able to correctly answer the question. If a judge answered a question incorrectly they were more likely to change their rating to be closer in line with the conditional p-value provided; conversely, if they answered the question correctly they were unlikely to make much of a change at all. These results hold for both the Minimally Qualified SMEs and the Highly Qualified SMEs. Furthermore, the Minimally Qualified SMEs made larger changes in their ratings than the Highly Qualified SMEs. This suggests that the ability level of the subject matter expert does affect the ratings provided. If the judges were not able to answer the questions then they relied more heavily on the conditional p-value feedback. Therefore, the selection criteria for judges should consider whether they are sufficiently able to answer the questions correctly in order to provide informative ratings.

One of the primary criticisms of the Angoff method is that judges are unable to accurately estimate the difficulty of items for minimally qualified candidates (Busch & Jaeger, 1990; Clauser et al., 2002; Cross et al., 1984; Impara & Plake, 1998; Reckase, 2000). This study finds that although the correlation of judge ratings to conditional p-values before the introduction of feedback was not high, the introduction of feedback did not increase the correlation to a practically significant degree; the judges seemed relatively confident in their initial ratings. However, I would argue that this finding does not support the view of Shepard et al. (1993) in determining that the Angoff method is fundamentally flawed. Rather, I contend that the issue is more that of judge selection criteria and ensuring that those participating in standard setting exercise be appropriately qualified and able to correctly answer the questions. As previously noted, the Highly

Qualified SMEs had a better initial sense of item difficulty, which just speaks to the adage that "you don't know what you don't know." The Minimally Qualified SMEs simply had difficulty in estimating their own ability.

An additional criticism is that the judges are too reliant on the feedback provided (Hurtz & Auerbach, 2003; Maurer & Alexander, 1992; Truxillo et al., 1996). The results here indicate that this is clearly not the case. Although the correlations of ratings to conditional p-value clearly increase following the introduction of feedback, it hardly seems that the judges are relying on the feedback. Judges are typically instructed to incorporate the feedback as a supplement to their opinion as a content expert. It seems that this is exactly what they're doing. However, this study had the luxury of a large sample size and lack of group effect. In a group setting the feedback may serve as a convenient point upon which the judges may converge, but that is an issue of group effect more than being overly reliant on the feedback.

Previous studies have found that inter-rater agreement increased between rounds following rater discussion, but this discussion did not increase the correlation between ratings and conditional p-values. The correlation between ratings and conditional p-values did not increase until the introduction of some form of empirical item-level feedback. The ability of a group of judges to come to a common consensus regarding item difficulty is often seen as one of the benefits of the Angoff method; however, it is also possible that a strong personality in a group could sway the ratings. Clauser et al. (2002) found a substantial group effect and noted that discussion without feedback improved judge agreement within groups, but not between groups. The inability of

groups of judges to provide consistent results across groups is one of the primary criticisms of the Angoff method and led to Clauser et al. (2014) and Hambleton et al. (2012) recommending that standard setting panels be conducted with multiple groups. This study was conducted asynchronously, eliminating the confounding inter-rater effect and allowing for an analysis of the perceptions of each individual judge.

Another notable difference in this study is that the feedback followed each item rather than being provided between rounds. Typically, judges rate all items, hold a discussion, examine feedback, and then provide a final rating. The methodology utilized here whereby judges provided an initial rating, received feedback, and then provided a final rating on an item before moving onto the next item allows for an analysis of each item independent of the other items. It may also allow judges to calibrate their internal sense of item difficulty with the conditional p-values early in the exercise rather than having to recalibrate between rounds.

Some researchers (Cizek, 1996; Hambleton & Pitoniak, 2006; Kane, 1994; Loomis, 2012; Raymond & Reid, 2001) have proposed that all relevant stakeholders for an examination should be invited to participate as judges in standard setting exercises. This study demonstrates that inviting individuals who are not content experts, and likely do not have the ability to correctly answer the questions, would negatively affect the results of a standard setting exercise and the resulting recommended cut score. For example, a member of the general public invited to participate in a standard setting study for medical licensure would lean so heavily on the feedback provided that it would be of little sense to have them involved. In this same scenario, if the feedback provided was a

conditional p-value based on the ability of a minimally qualified candidate, recalling that the calculation to determine the ability of a minimally qualified candidate is based on the current passing standard, the recommended passing score for this judge would be a self-fulfilling prophecy of retaining the current standard. Furthermore, the results here suggest that not only should non-content experts be excluded from participating, but also that those who have recently passed an exam may not be sufficiently qualified to participate in setting passing standards.

To my knowledge, there are no published research findings on standard setting exercises that utilize an asynchronous design. Harvey and Way (1999) and Harvey (2000) discuss the creation of a web-based application to conduct standard setting exercises and the differences in how judges felt about their experience using the web-based application compared to an in-person session, while MacCann and Stanley (2010) outline some of the potential benefits of a web-based standard setting exercise. The design utilized here allows for an examination of the effects of feedback on individual judges without the influence of a group effect. Future standard setting studies, both operational and research-based, should consider utilizing a similar design. The ability to eliminate the group effect and isolate the ratings and subsequent recommended cut score of an individual judge, while at the same time significantly increasing the number of judges involved, should lend itself to an increase in the reliability and validity of standard setting exercises. However, there clearly needs to be additional research conducted in order to support this claim.

The results of this study contribute to the body of evidence on the effects of feedback on ratings and the criteria for judge selection. The outcome of high-stakes testing determines whether an individual receives a diploma, gets into certain colleges and universities, and is granted entry into certain professions. In medical licensure, the outcome can literally be a matter of life or death if an unqualified physician is granted license to perform certain procedures. Cizek (2012) notes that the determination of cut scores also influences decisions about whether or not death penalty sentences should be carried out. With such broad and sweeping consequences, it is incumbent upon those conducting standard setting exercises to utilize the most rigorous methods available. In his 2004 review of the literature, Paul Brandon concludes that those conducting standard setting research are "not attending to the most rudimentary prescriptions about describing methods in sufficient detail to evaluate or replicate standard setting studies" (p.80). Furthermore, Brandon rues the "lack of a comprehensive program of standard setting research" (p.80). Clearly, research in this field continues to be necessary if we are ever to come to a consensus on the appropriate methods for setting passing standards.

Conclusions

Limitations

The primary limitation of this study is that it is correlational research and, while useful to help uncover the relationship between variables, does not provide and conclusive evidence for causation and often leads to more questions than answers. There is also an issue of whether the results of this study are generalizable since the judges were all board certified physicians and the empirical item feedback may affect these highly

trained content experts differently than content experts in other fields. Additionally, the relatively small size of the Minimally Qualified SME group is an issue. A larger sample of Minimally Qualified SMEs may have provided a more nuanced look at how the ability level of the SMEs influences the ratings provided.

A final limitation is that this standard setting exercise was conducted asynchronously and not in-person as is customary in Angoff-style standard setting. However, I see this as a benefit rather than a limitation. The ability of a group of judges to come to a common consensus regarding item difficulty is often seen as one of the benefits of the Angoff method. For this study there is no inter-judge agreement and the ability of one strong personality to sway the opinion of the group is not an issue. The asynchronous nature of this study also allows for the analysis of each individual judge without the additional confounding variable of judge inter-judge agreement. There is a fundamental difference in the process using an asynchronous method from the traditional in-person method. The in-person method focuses on discussion and consensus building, while the asynchronous method focuses on each judge constructing their own individual idea of a minimally qualified candidate. Although the outcome of the process is the same, recommending a minimum passing standard, this difference is critical to the analysis of the feedback mechanism and investigating how judges come to understand the idea of a minimally qualified candidate.

Educational Significance

The modified-Angoff method remains one of the most popular methods for setting performance standards in testing and assessment (Hurtz & Auerbach, 2003; Plake

& Cizek, 2012). It is common practice to provide content experts with feedback regarding the item-level difficulties; however, it is unclear how this feedback affects the ratings and recommendations of content experts. Recent research seems to indicate mixed results, noting that the feedback given to judges may or may not alter their ratings depending on the type of data provided, when the data was provided, and how judges collaborated within groups and between groups. The research proposed here seeks to examine issues related to the effects of item-level feedback on the ratings provided by judges. The results of this research may hold implications for how standard setting studies are conducted with regard to the difficulty and ordering of items, the ability level of content experts invited to participate in these studies, and the type of feedback that is provided to judges.

Standard setting methods that utilize judges and ask them to conceptualize a minimally qualified candidate (e.g. Angoff, Bookmark, etc.) will always have an issue regarding variance among what judges consider to be minimally qualified. This will be true in the realm of medical testing, K-12 testing, or any other field that utilized standardized tests. In high-stakes testing, setting performance standards is of critical importance and it is imperative that the utmost care be taken to ensure that standard setting exercises are conducted with the strongest theoretical and empirical foundation possible. I hope this research will add to the body of literature regarding the way in which standard setting studies are conducted and that researchers will begin to form a consensus around best-practices in setting performance standards.

Chapter Five

Effects of Incorrect Feedback on Judges' Ratings

Introduction

Setting performance standards is a judgmental process involving human opinions and values as well as technical and empirical considerations. Although all cut score decisions are by nature arbitrary, they should not be capricious (AERA et al., 2009; Cizek, 2012; Shepard, 1979). Establishing a minimum passing standard is the technical expression of a policy decision. The information gained through standard setting studies informs these policy decisions. To this end, it is necessary to conduct robust examinations of standard setting studies in order to understand how the information gained from standard setting studies influences policy decisions.

Examining how information regarding item-level feedback influences the perceptions of item difficulty held by content experts is a subject that has not been studied extensively. However, the way in which information regarding item-level feedback influences content experts' decisions may hold extensive consequences with regard to setting an appropriate passing standard and the subsequent pass/fail or other categorical decisions. In particular, Hambleton et al. (2012) call for more research on the empirical results of performance data, specifically noting the interesting questions raised by providing the incorrect data in Clauser, Mee, et al. (2009). The current study seeks to examine how the item-level feedback provided to content experts affects the ratings they provide.

The primary research questions guiding this study are:

1. How does the item-level feedback provided to content experts influence the ratings they provide?
2. Does altering the feedback given to content experts affect the ratings they provide?

Participant Selection

After obtaining assent from all participating member boards, an email was sent to all eligible individuals requesting their participation in a standard setting exercise. Invitations (n=228) were sent to those who were certificate holders in good standing who passed the 2011 or 2012 certification exam with a score of 540 or higher. Of the 228 invitations, 57 individuals accepted and completed the training, and 49 ratings. Following data cleaning, 46 were fit for use.

Participant Training

All volunteers were required to complete a web-based training session of approximately 30-45 minutes in length. The group sessions were typically conducted over the course of a week and at varying times to account for volunteers in different time zones. Individual sessions were also available for those who were unable to participate in a group session. Additional assistance and technical support was available throughout the process by phone and email.

The primary focus of the training sessions was to familiarize judges with the modified-Angoff method and discuss a key concept of this method, that of a "minimally knowledgeable, yet certifiable candidate". The modified-Angoff method asks judges to determine the probability of a minimally-competent candidate answering a question correctly. Judges were asked to think of a physician they knew who they believe lacks the knowledge sufficient to be a board certified physician. They were then asked to think of a physician they knew who they believe would be considered barely qualified to be a board certified physician. In order to help conceptualize their understanding of a minimally qualified candidate, statements such as, "…this person would not be highly knowledgeable, but you would still be comfortable with them receiving the same certification that you have" were presented for consideration.

Judges were also provided an overview of the web-based rating software, including screenshots and instructions for accessing the website. Following a brief discussion on the mechanics of using the software, participants were provided an explanation of the concept of conditional p-values, or the percentage of minimally-qualified candidates who would answer a question correctly.

Finally, judges were shown a copy of a survey that would be administered once they completed rating each of the 120 items. Each survey question was reviewed to ensure that judges had a clear understanding of what was being asked and an understanding of how these questions factored into the standard setting process. This research was conducted during an operational standard setting exercise. The operational standard setting exercise utilized three distinct standard setting methods: (1) a modified-

Angoff method, (2) the Hofstee Method (1983), and (3) the Beuk Compromise Adjustment (1984). The responses to the survey questions are critical to the implementation of the Hofstee and Beuk methods, but are not within the scope of this research.

Data Collection

Item rating process

The asynchronous item rating process was designed to maximize participation by allowing judges to enter their ratings at their convenience during the rating window. This was accomplished through the use of a web-based software application that was available to the judges 24-hours a day during the rating window. For this study the rating window was open for 18 days. The asynchronous nature of the process also eliminated the inconvenience and expense associated with requiring judges to travel. Although volunteers were not reimbursed for their time, they were recognized for their contributions with a framed acknowledgement.

The item rating process consisted of judges providing multiple data entries for 120 individual items. Judges begin by attempting to answer the question correctly and providing an initial difficulty rating for that question using a scale from 0-100. During training the judges were informed that this rating is the percentage of minimally qualified candidates that they believe would get this question correct. Thus, 0 would be a difficult question and 100 would be an easy question. Once an answer and initial rating have been locked-in, judges are provided with the correct answer to the question and the conditional

p-value, which is a calculation of the percentage of examinees with ability levels at the current passing standard that would get that question correct. The conditional p-value is provided on the same 0-100 scale that judges use to rate the items. After receiving feedback regarding the correct answer and conditional p-values, judges are able to adjust their ratings for the item and submit a final rating. Judges are also asked a multiple choice question regarding their perception of the item and allowed to provide comments, but these issues are beyond the scope of this study.

Variables and Data Elements

The data returned contained 8 data elements, of which 3 are outside the scope of this research. The pertinent data elements are UserID, Form, Initial Rating, Final Rating, and Response Vector. The UserID variable is a unique identifier assigned to each judge that allows the responses collected from the standard setting software to be matched with the associated demographic information, which will be discussed shortly. The Form variable indicates which version of the standard setting items the judge saw. For this study all judges saw the same form. The Initial Rating, Final Rating, and Response Vector variables exist for every item and are labeled sequentially according to the item sequence. For example, the Initial Rating variables in the dataset are labeled "InitialRating_1", "InitialRating_2", " InitialRating_3"...etc. Therefore, there exist initial ratings, final ratings, and response vectors for each item rated by each judge. The Initial Rating variable is the 0-100 rating provided by each judge on each item before they received feedback. The Final Rating variable is the 0-100 rating provided by each judge on each item after they received feedback. The Response Vector variable is a 0-1 scoring

for every item based on whether the judge answered the question correctly (1) or incorrectly (0) before they received any feedback.

As previously mentioned, the UserID variable allows for matching rating sets to the appropriate demographic information. The demographic information for each judge included gender, medical degree (i.e. MD or DO), and score on the last exam. This demographic information is used to ensure that there is adequate representation and that judge selection is not biased.

Data Cleaning

As is typically the case, the results from the item rating process were returned with missing data points as well as instances of misuse of the rating scale by judges. The most common issue was that of misuse of the rating scale. Judges often provided single-digit ratings of item difficulty. On the 0-100 scale, a single-digit response would represent a question so difficult that less than 10 percent of minimally qualified examinees would answer it correctly. Judges often included comments indicating that they had made some kind of mistake in providing the rating for specific items. In other cases it was relatively obvious that a typographical error existed. In each of these cases, the single-digit ratings were transformed onto the 100-point scale by multiplying the rating provided by 10. Although there is no rule governing the extent to which these transformations were tolerated, if a judge had multiple instances (typically more than 10) they were removed from the dataset due to significant misuse of the scale. It was also common for judges to provide ratings for only a portion of the items. If any number of the ratings were missing that judge was removed for having an incomplete dataset.

For this study, one judge provided the first three initial ratings and first two final ratings using a 10-point scale, but then corrected to using the 100-point scale. Another judge provided three initial ratings and three final ratings using the 10-point scale. A third judge provided a single initial rating and final rating using the 10-point scale. Two judges provided an initial rating using the 10-point scale, but corrected the final rating. In each of these instances, the single-digit scores were transformed onto the 100 point scale by multiplying the rating provided by 10.

Three judges were removed from this dataset. One judge provided only the first eleven ratings, while another provided ratings for 100 of the 120 items. Both of these judges were removed for providing incomplete datasets. A final judge provided all 120 items on a 10-point scale. This judge was removed for incorrect use of the scale.

Creating conditional p-values

The modified -Angoff method utilized here is a content-based method for recommending a passing standard that asks content experts to examine each item and determine the probability of a "minimally competent examinee" answering a question correctly. Judges are commonly provided some form of past examinee performance data to assist in their decision-making process. Each of the examinations under consideration here was scored using the dichotomous Rasch model (Rasch, 1960). The Rasch model is a logistic model of latent traits that provides person measures and item difficulties in log-odds units, commonly referred to as logits. There are two options available for providing performance data to judges regarding the probability of a minimally competent candidate answering a question correctly: (1) aggregate the item-level responses only from previous

examinees who scored at the passing standard, and (2) calculate conditional p-values for each item. Using the responses from previous examinees may provide a more accurate reflection of the ability of minimally competent examinees, but there is typically a dearth of examinees who scored at the passing standard for each item under examination. Therefore, the typical action is to create conditional p-values for each item. Conditional p-values are calculations of the percentage of candidates with ability estimates at the passing standard expected to get the question correct.

Providing judges a calculation of conditional p-values rather than overall item difficulty is done in order to give standard setting judges a more accurate view of how minimally competent examinees would actually perform on this item rather than relying on judges trying to estimate a probability of success for each item based on their own sense of how a minimally competent examinee might perform. However, the overall item difficulty is used to calculate the conditional p-values. The transformation of the overall item difficulty into a conditional p-value was accomplished using the following formula:

$$\Pr\{\beta_{MPS} = 1\} = \frac{e^{MPS-\delta i}}{1 + e^{MPS-\delta i}}$$

Where:

$\Pr\{\beta_{MPS} = 1\}$	is the probability of a correct response by a minimally qualified candidate
MPS	is the calibration of the minimum passing standard
δi	is the difficulty of item i
e	is the base of the natural logarithm

The overall item difficulty calibration in logits from previous exam administrations is subtracted from the calibration in logits for the minimum passing score to produce a conditional difficulty measure of each item for a minimally qualified candidate. The base of the natural logarithm for this new calibration is divided by 1 + the base of the natural logarithm, which produces a conditional probability. This conditional probability is multiplied by 100 in order to return a percentage in a whole number that judges can readily understand. This percentage is referred to as the conditional p-value of an item.

For this study, the following formula was used to derive the conditional p-values:

$$\Pr\{\beta_{MPS} = 1\} = \frac{e^{\delta i - MPS}}{1 + e^{\delta i - MPS}}$$

Which gives the inverse of the conditional p-value, such that an item with a correct conditional p-value of 90 would be shown to have a conditional p-value of 10; a correct conditional p-value of 60 would be shown as 40; and a correct conditional p-value of 50 would still be shown to be 50.

Methods

The primary research question for this study is, "How does item-level feedback provided to content experts influence the ratings they provide?" In order to fully explore this question, I employ several strategies designed to examine the various ways in which the final ratings of judges are influenced by the feedback provided. There were two types of feedback provided to judges in this standard setting exercise: (1) conditional p-values of item difficulty, and (2) whether the judge was able to correctly answer the question.

However, a primary concern of this study was that judges were provided an inverted conditional p-value of item difficulty. So, an item that 10 percent of minimally qualified candidates would be expected to get correct was said to have been answered correctly by 90 percent of minimally qualified candidates. Therefore, in addition to the primary research question investigating item-level feedback, a second research question asks, "does altering the feedback given to content experts affect the ratings they provide?"

In order to examine these two research questions I will perform a one-way repeated measures ANOVA to determine whether the means for the initial and final ratings are significantly different. Next, in order to examine the effects of the conditional p-value feedback, I will calculate a Pearson product-moment correlation coefficient for each judge's initial ratings with the conditional p-values and their final ratings with the conditional p-values. A stronger correlation coefficient would suggest that the judge adjusted their rating to be more in line with the conditional p-values. Finally, in order to examine the effect of whether the judge answered the question correctly, I will examine their average change in ratings for questions answered correctly and as well as the average change for those questions answered incorrectly. A paired-samples t-test will be employed to determine whether the mean change for correct answers is significantly different from the mean change for incorrect answers.

It is also important to examine whether there are differences in the ratings provided based on demographic variables. Therefore, repeated measures factorial (mixed) ANOVA tests will be conducted to determine whether there are any interaction effects based on certain demographic variables. For this analysis the only demographic

variable with enough statistical power was gender. There was not enough diversity between medical degree (MD, DO) and participating board to provide useful information.

Results

The results of the one-way ANOVA show that the judges' ratings were not significantly affected by the feedback provided, $F(1, 45) = 1.74$, $p=1.94$. With regard to the effects of the conditional p-value feedback, the correlation of the judges' ratings with the conditional p-values provided increased following the introduction of the feedback (Table 5.1). Of the 46 judges, 1 (2.2%) did not change their ratings to a degree that it altered their correlation coefficient, 6 (13.0%) changed their ratings such that their correlation coefficient increased by less than .1, 36 (78.3%) changed their ratings such that their correlation coefficient increased by .1 or more, and 3 (6.5%) changed their ratings such that their correlation coefficient decreased. Furthermore, of the 36 whose correlation coefficient increased by more than .1, there were 2 whose correlation coefficients increased by more than 1.0; changing from a negative correlation to a strongly positive correlation. These two judges relied almost entirely on the conditional p-values provided to make their final ratings.

Table 5.1.
Summary of Correlation with Conditional P-value feedback

	Initial Rating		Final Rating	
	Mean	SD	Mean	SD
Incorrect P-value	-0.10	0.13	0.21	0.32
Correct P-value	-0.08	0.14	-0.20	0.17

Although the final ratings were more strongly correlated with the conditional p-value feedback provided, the initial ratings remained associated with their final ratings as seen in Figure 5.1. This figure shows each judge's initial rating (before feedback is provided) on the X-axis plotted against their final rating (following feedback) on the Y-axis for each item. A linear regression line and associated R-square value are also provided.

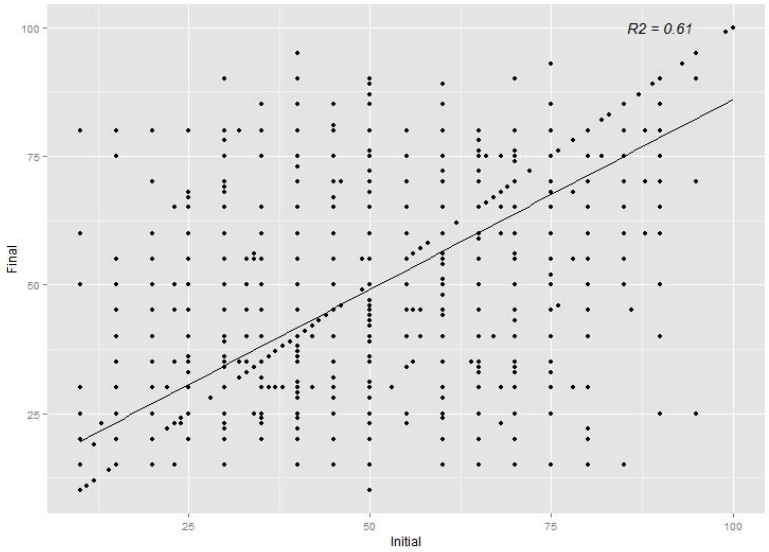

Figure 5.1. Plot of Initial Rating with Final Rating.

With regard to the effect of judges answering the question correctly, on average the judges changed their rating to a significantly greater degree when answering a question incorrectly (M=6.7, SE =.66) as opposed to answering a question correctly (M=5.5, SE=.66), t (45) = -2.44, p <.05. An examination of the interaction effects of demographic variables showed there was no significant interaction effect of gender, indicating that the ratings provided by male and female judges were generally the same, F(1, 44) = .85, p=.36.

Discussion

This study sought to explore how item-level feedback provided to content experts affected the ratings they provide and how providing incorrect conditional p-value feedback affected the ratings. The results indicate that judges did not utilize the incorrect conditional p-value feedback. It is interesting to note that six judges (13%) had final correlation coefficients greater than .6 (.94, .85, .85, .75, .69, & .65). These judges relied heavily on the feedback provided, ignoring their initial sense of item difficulty. However, the rest of the judges ignored the feedback almost entirely. Table 5.1 shows that the initial ratings has a slightly stronger correlation with the correct conditional p-values, but following feedback the judges altered their ratings to be more in line with the incorrect conditional p-values that were provided as feedback. This is a bit of an anomaly and appears to be driven by those 6 judges that altered their ratings to a high degree.

Although the judges did not seem to use the conditional p-value feedback, whether or not they answered the question correctly did have an effect. Figure 5.1 shows that the association between initial ratings and final ratings remained strong even after feedback, which suggests that judges tend to primarily rely on their innate sense of item difficulty rather than the conditional p-values provided. However, if a judge answered a question incorrectly they were more likely to change their rating to be closer in line with the conditional p-value provided; conversely, if they answered the question correctly they were unlikely to make much of a change at all.

These results suggest that the impact of providing incorrect conditional p-value feedback varies wildly by judge. Some judges completely ignored the feedback and relied on their own sense of item difficulty and other completely changed their ratings to mimic the feedback. If conditional p-value feedback is to be provided to standard setting judges, it is of critical importance that it be accurate and properly interpreted by the judges. If not, the results of a standard setting exercise will be so disparate as to be completely unintelligible.

Two previous studies investigated the effects of manipulating the feedback provided to judges. Clauser, Mee, et al. (2009) found that judges incorporated the feedback whether it was correct or not and concluded that judges relied on the data when discrepancies between their expectation and the data were present. Mee et al. (2013) informed the judges that some of the data was incorrect and then found that the judges utilized the feedback less than in a parallel study – this result is hardly surprising. The current study adds to this body of literature by suggesting that the size of the discrepancy should be considered. Clauser, Mee, et al. (2009) and Mee et al. (2013) utilized relatively small changes to the feedback, while this study completely inverted the conditional p-values. When faced with large discrepancies, the judged did not substantially utilize the feedback, placing limits on the findings of Clauser, Mee, et al. (2009) and Mee et al. (2013). Furthermore, the results here suggest that the discrepancies that drove the findings of previous research may have been the ability of content experts to correctly answer the questions; an issue not addressed in those studies.

One of the primary criticisms of the Angoff method is that judges are unable to accurately estimate the difficulty of items for minimally qualified candidates (Busch & Jaeger, 1990; Clauser et al., 2002; Cross et al., 1984; Impara & Plake, 1998; Reckase, 2000). Table 5.1 would seem to support the conclusion of Shepard et al. (1993) in determining that the Angoff method is fundamentally flawed. However, the first item was a very hard item with a conditional p-value of 93 and judges were told it had a conditional p-value of 7. The second item was an easy item with a conditional p-value of 80 and judges were told it had a conditional p-value of 20. So, right from the start the judges' sense of item difficulty was attacked. Any attempt to draw conclusions about their ability to accurately estimate the difficulty of items for minimally qualified candidates should be taken with a grain of salt. I would argue that since judges who answered the question incorrectly were more likely to utilize the feedback than those who answered correctly, a judge's ability to correctly answer a question overrides their professional sense of item difficulty. This would seem to be more of an issue regarding judge selection criteria and ensuring that those participating in standard setting exercise be appropriately qualified and able to correctly answer the questions.

An additional criticism is that the judges are too reliant on the feedback provided (Hurtz & Auerbach, 2003; Maurer & Alexander, 1992; Truxillo et al., 1996). The results here indicate that this is clearly not the case. For the most part, judges recognized the disconnect between the incorrect feedback and their professional sense of item difficulty and ignored the incorrect feedback almost entirely. Judges are typically instructed to incorporate the feedback as a supplement to their opinion as a content expert. It seems that this is exactly what they're doing. However, this study had the luxury of a large

sample size and lack of group effect. In a group setting the feedback may serve as a convenient point upon which the judges may converge, but that is an issue of group effect more than being overly reliant on the feedback.

Previous studies have found that inter-rater agreement increased between rounds following rater discussion, but this discussion did not increase the correlation between ratings and conditional p-values. The correlation between ratings and conditional p-values did not increase until the introduction of some form of empirical item-level feedback. The ability of a group of judges to come to a common consensus regarding item difficulty is often seen as one of the benefits of the Angoff method; however, it is also possible that a strong personality in a group could sway the ratings. Clauser et al. (2002) found a substantial group effect and noted that discussion without feedback improved judge agreement within groups, but not between groups. The inability of groups of judges to provide consistent results across groups is one of the primary criticisms of the Angoff method and led to Clauser et al. (2014) and Hambleton et al. (2012) recommending that standard setting panels be conducted with multiple groups. This study was conducted asynchronously, eliminating the confounding inter-rater effect and allowing for an analysis of the perceptions of each individual judge.

Another notable difference in this study is that the feedback followed each item rather than being provided between rounds. Typically, judges rate all items, hold a discussion, examine feedback, and then provide a final rating. The methodology utilized here whereby judges provided an initial rating, received feedback, and then provided a final rating on an item before moving onto the next item allows for an analysis of each

item independent of the other items. It may also allow judges to calibrate their internal sense of item difficulty with the conditional p-values early in the exercise rather than having to recalibrate between rounds. Typically, calibrating early would be a benefit for the judge; however, as previously mentioned, for this study the first item was a very hard item with a conditional p-value of 93 and judges were told it had a conditional p-value of 7. The second item was an easy item with a conditional p-value of 80 and judges were told it had a conditional p-value of 20. So, right from the start this study attempted to significantly alter the judges' sense of item difficulty and they were likely skeptical of the feedback provided for the entire set of items.

Some researchers (Cizek, 1996; Hambleton & Pitoniak, 2006; Kane, 1994; Loomis, 2012; Raymond & Reid, 2001) have proposed that all relevant stakeholders for an examination should be invited to participate as judges in standard setting exercises. This study demonstrates that inviting individuals who are not content experts, and likely do not have the ability to correctly answer the questions, would negatively affect the results of a standard setting exercise and the resulting recommended cut score. For example, a member of the general public invited to participate in a standard setting study for medical licensure would lean so heavily on the feedback provided that it would be of little sense to have them involved. In this same scenario, if the feedback provided was a conditional p-value based on the ability of a minimally qualified candidate, recalling that the calculation to determine the ability of a minimally qualified candidate is based on the current passing standard, the recommended passing score for this judge would be a self-fulfilling prophecy of retaining the current standard. Furthermore, the ability of a content expert to sense a disconnect between their professional sense of item difficulty and the

feedback provided is precisely the reason that content experts are be utilized for setting performance standards.

To my knowledge, there are no published research findings on standard setting exercises that utilize an asynchronous design. Harvey and Way (1999) and Harvey (2000) discuss the creation of a web-based application to conduct standard setting exercises and the differences in how judges felt about their experience using the web-based application compared to an in-person session, while MacCann and Stanley (2010) outline some of the potential benefits of a web-based standard setting exercise. The design utilized here allows for an examination of the effects of feedback on individual judges without the influence of a group effect. Future standard setting studies, both operational and research-based, should consider utilizing a similar design. The ability to eliminate the group effect and isolate the ratings and subsequent recommended cut score of an individual judge, while at the same time significantly increasing the number of judges involved, should lend itself to an increase in the reliability and validity of standard setting exercises. However, there clearly needs to be additional research conducted in order to support this claim.

The results of this study contribute to the body of evidence on the effects of feedback on ratings and the criteria for judge selection. The outcome of high-stakes testing determines whether an individual receives a diploma, gets into certain colleges and universities, and is granted entry into certain professions. In medical licensure, the outcome can literally be a matter of life or death if an unqualified physician is granted license to perform certain procedures. Cizek (2012) notes that the determination of cut

scores also influences decisions about whether or not death penalty sentences should be carried out. With such broad and sweeping consequences, it is incumbent upon those conducting standard setting exercises to utilize the most rigorous methods available. In his 2004 review of the literature, Paul Brandon concludes that those conducting standard setting research are "not attending to the most rudimentary prescriptions about describing methods in sufficient detail to evaluate or replicate standard setting studies" (p.80). Furthermore, Brandon rues the "lack of a comprehensive program of standard setting research" (p.80). Clearly, research in this field continues to be necessary if we are ever to come to a consensus on the appropriate methods for setting passing standards.

Conclusions

Limitations

The primary limitation of this study is that it is correlational research and, while useful to help uncover the relationship between variables, does not provide and conclusive evidence for causation and often leads to more questions than answers. There is also an issue of whether the results of this study are generalizable since the judges were all board certified physicians and the empirical item feedback may affect these highly trained content experts differently than content experts in other fields. Additionally, this was not a true experimental design. Ideally, there would have been a control group and an experimental group, which would have provided for a more robust examination of the differences. However, there is no reason that we should not be able to learn something from mistakes and the results of this study suggest that future research should be conducted in this subject.

A final limitation is that this standard setting exercise was conducted asynchronously and not in-person as is customary in Angoff-style standard setting. However, I see this as a benefit rather than a limitation. The ability of a group of judges to come to a common consensus regarding item difficulty is often seen as one of the benefits of the Angoff method. For this study there is no inter-judge agreement and the ability of one strong personality to sway the opinion of the group is not an issue. The asynchronous nature of this study also allows for the analysis of each individual judge without the additional confounding variable of judge inter-judge agreement. There is a fundamental difference in the process using an asynchronous method from the traditional in-person method. The in-person method focuses on discussion and consensus building, while the asynchronous method focuses on each judge constructing their own individual idea of a minimally qualified candidate. Although the outcome of the process is the same, recommending a minimum passing standard, this difference is critical to the analysis of the feedback mechanism and investigating how judges come to understand the idea of a minimally qualified candidate.

Educational Significance

The modified-Angoff method remains one of the most popular methods for setting performance standards in testing and assessment (Hurtz & Auerbach, 2003; Plake & Cizek, 2012). It is common practice to provide content experts with feedback regarding the empirical item difficulties; however, it is unclear how this feedback affects the ratings and recommendations of content experts. Recent research seems to indicate mixed results, noting that the feedback given to judges may or may not alter their ratings

depending on the type of data provided, when the data was provided, and how judges collaborated within groups and between groups. The research proposed here seeks to examine issues related to the effects of item-level feedback on the ratings provided by judges. The results of this research may hold implications for how standard setting studies are conducted with regard to the difficulty and ordering of items, the ability level of content experts invited to participate in these studies, and the type of feedback that is provided to judges.

Standard setting methods that utilize judges and ask them to conceptualize a minimally qualified candidate (e.g. Angoff, Bookmark, etc.) will always have an issue regarding variance among what judges consider to be minimally qualified. This will be true in the realm of medical testing, K-12 testing, or any other field that utilized standardized tests. In high-stakes testing, setting performance standards is of critical importance and it is imperative that the utmost care be taken to ensure that standard setting exercises are conducted with the strongest theoretical and empirical foundation possible. I hope this research will add to the body of literature regarding the way in which standard setting studies are conducted and that researchers will begin to form a consensus around best-practices in setting performance standards.

References

AERA, APA, & NCME. (2009). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Board of Medical Specialties. (2014). The Specialty Board Movement. Retrieved October 2, 2014, from http://www.abms.org/About_ABMS/ABMS_History/Extended_History/Specialty_Board_Movement.aspx

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

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*(2), 147-152. doi: 10.1111/j.1745-3984.1984.tb00226.x

Brandon, P. R. (2004). Conclusions About Frequently Studied Modified Angoff Standard-Setting Topics. *Applied Measurement in Education, 17*(1), 59-88.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*(2), 145-163. doi: 10.1111/j.1745-3984.1990.tb00739.x

Cizek, G. J. (1993). Reactions to National Academy of Education report "Setting performance standards for student achievement". Washington, DC: National Assessment Governing Board.

Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues & Practice, 15*(1), 13-21.

Cizek, G. J. (2012). An Introduction to Contemporary Standard Setting: Concepts, Characteristics, and Contexts. In G. J. Cizek (Ed.), *Setting Performance Standards* (2nd ed., pp. 3-14). New York: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A guide to establishing and evaluating performance standards on tests* Thousand Oaks, CA: Sage.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). An Empirical Examination of the Impact of Group Discussion and Examinee Performance Information on Judgments Made in the Angoff Standard-Setting Procedure. *Applied Measurement in Education, 22*(1), 1-21. doi: 10.1080/08957340802558318

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement, 46*(4), 390-407.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The Effect of Data Format on Integration of Performance Data into Angoff Judgments. *International Journal of Testing, 13*(1), 65-85.

Clauser, B. E., Swanson, D. B., & Harik, P. (2002). Multivariate Generalizability Analysis of the Impact of Training and Examinee Performance Information on Judgments Made in an Angoff-Style Standard-Setting Procedure. *Journal of Educational Measurement, 39*(4), 269-290. doi: 10.2307/1435404

Clauser, J. C., Margolis, M. J., & Clauser, B. E. (2014). An Examination of the Replicability of Angoff Standard Setting Results Within a Generalizablility Theory Framework. *Journal of Educational Measurement, 51*(2), 127-140.

Cordes, F. C., & Rucker, C. W. (1961). History of the American Board of Ophthalmology. *Trans Am Ophthalmol Soc, 59*, 295-332.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National teacher examinations. *Journal of Educational Measurement, 21,* 113-129. doi: 10.1111/j.1745-3984.1984.tb00224.x

Goodwin, L. D. (1999). Relations Between Observed Item Difficulty Levels and Angoff Minimum Passing Levels for a Group. *Applied Measurement in Education, 12*(1), 13.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 83-116). Nahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., . . . Zwick, R. (2000). A Response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Science'Grading the Nations Report Card. *Educational Measurement: Issues and Practice, 19*(2), 5-14. doi: 10.1111/j.1745-3992.2000.tb00024.x

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433-470). Washington, DC: American Council on Education.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential Steps in Setting Performance Standards on Educational Tests and Strategies for Assessing teh Reliability of Results. In G. J. Cizek (Ed.), *Setting Performance Standards:* (2nd ed., pp. 47-76). New York: Routledge.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. New York: Sage Publications.

Harvey, A. L. (2000). *Comparing Onsite and Online Standard Setting Methods for Multiple Levels of Standards*. Paper presented at the National Council on Measurement in Education, New Orleans.

Harvey, A. L., & Way, W. D. (1999). *A Comparison of Web-Based Standard Setting and Monitored Standard Setting*. Paper presented at the National Council on Measurement in Education, Montreal.

Hofstee, W. K. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109-127). San Francisco: Jossey Bass.

Hurtz, G. M., & Auerbach, M. A. (2003). A Meta-Analysis of the Effects of Modifications to the Angoff Method on Cutoff Scores and Judgment Consensus. *Educational & Psychological Measurement, 63*(4), 584-601. doi: 10.1177/0013164403251284

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69-81. doi: 10.1111/j.1745-3984.1998.tb00528.x

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

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461. doi: 10.2307/1170678

Livingston, S. A., & Zieky, M. J. (1982). *Passing Scores*. Princeton, NJ: Educational Testing Service.

Loomis, S. C. (2012). Selecting and Training Standard Setting Participants. In G. J. Cizek (Ed.), *Setting Performance Standards* (2nd ed., pp. 107-134). New York: Routledge.

Loomis, S. C., & Borque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Standard performance standards: Concepts, methods, and perspectives* (pp. 175-217). Mahwah, NJ: Lawrence Erlbaum.

MacCann, R. G., & Stanley, G. (2010). Extending participation in standard setting: an online judging proposal. *Educational Assessment, Evaluation and Accountability, 22*, 139–157.

Margolis, M. J., & Clauser, B. E. (2014). The Impact of Examinee Performance Information on Judges' Cut Scores in Modified Angoff Standard-Setting Exercises. *Educational Measurement: Issues & Practice, 33*(1), 15-22.

Maurer, T. J., & Alexander, R. A. (1992). Methods of improving employment test critical scores derived by judging test content. *Personnel Psychology, 45*, 727-762.

Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The Impact of Process Instructions on Judges' Use of Examinee Performance Data in Angoff Standard Setting Exercises. *Educational Measurement: Issues & Practice, 32*(3), 27-35. doi: 10.1111/emip.12013

Plake, B., Melican, G., & Mills, C. (1991). Factors influencing intrajudge consistency during standard setting. *Educational Measurement: Issues & Practice, 10*(2), 15-16.

Plake, B. S., & Cizek, G. J. (2012). Variations on a Theme: The Modified Angoff, Extended Angoff, and Yes/No Standard Setting Methods. In G. J. Cizek (Ed.), *Setting Performance Standards* (2nd ed., pp. 181-199). New York: Routledge.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark.: Danish Institute for Educational Research.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Erlbaum.

Reckase, M. (2000). The evloution of the NAEP achievement levels setting process: A summary of research and development efforts conducted by NAEP. Iowa City: ACT.

Reckase, M. (2001). Innovative methods for helping standard setting participants to perform their task: The role of feedback regarding consistency, accuracy, and

impact. In G. J. Cizek (Ed.), *Standard Setting: Concepts, methods, and perspectives* (pp. 159-173). Mahwah, NJ: Erlbaum Associates.

Reid, J. (1991). Training judges to provide standard-setting data. *Educational Measurement: Issues & Practice, 10*(23), 11-14.

Shaffer, R. N. (1991). *The history of the American Board of Ophthalmology, 1916-1991*. United States: R.N. Shaffer.

Shepard, L. (1979). Setting Standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education* (pp. 59-71). Washington, DC: National Council on Measurement in Education.

Shepard, L. (1995). Implications for standard setting of the National Academy of Educational Evaluation of the National Assessment of Educational Progress achievement levels *Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board and the National Center for Educational Statistics* (pp. 143-159). Washington, DC: U.S. Government Printing Office.

Shepard, L., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). Setting Performance Standards for Student Achievement. Stanford, CA: National Academy of Education.

Stone, G. E. (2004). Objective Standard Setting (or Truth in Advertising). In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 445-459). Maple Grove, MN.: JAM Press.

Stone, G. E. (2006). Whose Criterion Standard Is It Anyway? *Journal of Applied Measurement, 7*(2), 160-169.

Stone, G. E., Beltyukova, S., & Fox, C. M. (2008). Objective Standard Setting for Judge-Mediated Examinations. *International Journal of Testing, 8*(2), 180-196. doi: 10.1080/15305050802007083

Stone, G. E., Koskey, K. L. K., & Sondergeld, T. A. (2011). Comparing Construct Definition in the Angoff and Objective Standard Setting Models: Playing in a House of Cards Without a Full Deck. *Educational and Psychological Measurement, 71*(6), 942-962. doi: 10.1177/0013164410394338

Thorndike, R. L. (Ed.). (1971). *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting Cutoff Scores for Personnel Selection Tests: Issues. Illustrations, and Recommendations. *Human Performance, 9*(3), 275.

Young, A., Chaudhry, H. J., Rhyne, J., & Dugan, M. (2010). A Census of Actively Licensed Physicians in the United States, 2010. *Journal of Medical Regulation, 96*(4), 10-20.

Michael Peabody

EDUCATION

University of Kentucky

(2011) Master of Science in Education - Higher Education

A Critical Analysis of the Identification and Treatment of First-Generation College Students: A Social Capital Approach.

(2002) Bachelor of Arts, Classics

PROFESSIONAL EXPERIENCE

6/2013 – present Psychometrician

The American Board of Family Medicine

Office of the Vice President for Psychometric Services

Lexington, KY

11/2012 – 6/2013 Institutional Research Specialist

Kentucky Community and Technical College System

Office of Research and Policy Analysis

Lexington, KY

9/2011 – 11/2012 SACS Reaffirmation of Accreditation Project

University of Kentucky

Office of the Vice President for Institutional Effectiveness

Lexington, KY

PRIMARY RESEARCH INTERESTS

- Rasch measurement
- Testing, Licensure, and Certification
- Survey Research
- Program Evaluation and Performance Measurement
- Higher Education Organization and Policy

PUBLICATIONS (peer-reviewed)

O'Neill, T.R., **Peabody, M**., Blackburn, B., & Peterson, L. (2014). Creating an I-SOP Scale for Family Practice. *Journal of Applied Measurement*, 15(3), 227-239.

O'Neill, T.R., **Peabody, M.R**., Jin Bee Tan, R., & Du, Y. (2013). How much item drift is too much? *Rasch Measurement Transaction, 27*(3), 1423-1424.

Peabody, M. (2013). The identification and treatment of first-generation college students: a social capital approach. *Kentucky Journal of Higher Education Policy and Practice*. Available at: http://uknowledge.uky.edu/kjhepp/vol2/iss1/4/

Peabody, M. (2011). "Recognizing and Serving Low-Income Students in Higher Education (review)," *Kentucky Journal of Higher Education Policy and Practice, 1*(1), Article 1. Available at: http://uknowledge.uky.edu/kjhepp/vol1/iss1/1

Hutchens, K, Deffendall, M., & **Peabody, M**. (2011). "Supporting First Generation College Students," *Kentucky Journal of Higher Education Policy and Practice, 1*(1), Article 4. Available at: http://uknowledge.uky.edu/kjhepp/vol1/iss1/4

PUBLICATIONS (non peer-reviewed)

Peabody, M.R., O'Neill, T.R., & Puffer, J.C. (in-press). Who performs better on the MC-FP Exam: Initial Certifiers or Experienced Physicians? *Journal of the American Board of Family Medicine*.

Peabody, M.R., O'Neill, T.R., & Puffer, J.C. (in-press). Who performs better on the MC-FP Exam: Initial Certifiers or Experienced Physicians? *Annals of Family Medicine*.

O'Neill, T.R., **Peabody, M.R**., & Puffer, J.C. (2013). The ABFM Begins to Use Differential Item Functioning. *The Annals of Family Medicine, 11*(6), 578-579. doi: 10.1370/afm.1587

O'Neill, T.R., **Peabody, M.R**., & Puffer, J.C. (2013). The ABFM Begins to Use Differential Item Functioning. *The Journal of the American Board of Family Medicine*, *26*(6), 807-809. doi: 10.3122/jabfm.2013.06.130239

REVIEWS

Peabody, M. (2012). "Inside the College Gates: How Class and Culture Matter in Higher Education (review)," *Teachers College Record*: ID Number:16654. Available at: http://www.tcrecord.org/Content.asp?ContentId=16654

PRESENTATIONS

Peterson, L., Blackburn, B., **Peabody, M**., & O'Neill, T. (2014) *Family Physicians' Scope of Practice and American Board of Family Medicine Recertification Examination Performance.* Paper presented at the annual meeting of the North American Primary Care Research Group, New York, NY.

Peterson, L., Finnegan, S., **Peabody, M**., O'Neill, T., & Phillips, R. (2014) *Family Physicians' Scope of Practice Varies by Practice Organization and Rural/Urban Status.* Paper presented at the annual meeting of the North American Primary Care Research Group, New York, NY.

Peabody, M. (2014). *Stealing Thurstone's Crime Scale: A method for paired comparisons.* Presentation at the annual meeting of the Ohio River Valley Objective Measurement Seminar, Cincinnati, OH.

Peterson, L., **Peabody, M.**, & O'Neill, T. (2014). *The Scope of Practice for Primary Care (SP4PC) Scale: A Psychometric Scale to Measure Scope of Practice.* Poster presented at the annual meeting of Academy Health, San Diego, CA.

Peabody, M. (2011). *"Promising Practices" in the Retention of First-Generation College Students.* Paper presented at the College Personnel Association of Kentucky annual meeting, Lexington, KY.

ASSESSMENT & EVALUATION REPORTS

O'Neill, T.R., & **Peabody, M.R.** (2014). Sports Medicine Certificate of Added Qualification Summer Psychometric Report. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2014). MC-FP Examination, Differential Item Functioning Report. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2014). Maintenance of Certification - Family Physicians (MC-FP) Spring Psychometric Report. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2014). Maintenance of Certification - Family Physicians (MC-FP): Standard Setting Report. Lexington, KY: American Board of Family Medicine

Hagen, M., O'Neill, T.R., & **Peabody, M.R.** (2014). Do Diplomates Think that the MC-FP Examination Items Require Using an External Searchable Resource? Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2014). Sports Medicine Certificate of Added Qualification: Standard Setting Report. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2013). Differential Item Functioning Report: MC-FP Examination. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2013). Annual 2013 Sports Medicine Certificate of Added Qualification: Psychometric Report. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2013). Annual 2013 Maintenance of Certification - Family Physicians (MC-FP), Psychometric Report. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2013). In-Training Examination Score Results Handbook. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2013). Sports Medicine Certificate of Added Qualification Summer Psychometric Report. Lexington, KY: American Board of Family Medicine.

O'Neill, T.R., & **Peabody, M.R.** (2013). Maintenance of Certification - Family Physicians (MC-FP) Spring Psychometric Report. Lexington, KY: American Board of Family Medicine.

Peabody, M. (2013). KBEMS 2012 Renewal Survey. KCTCS Office of Research and Policy Analysis for the Kentucky Board of Emergency Medical Services. Versailles, KY: Kentucky Board of Emergency Medical Services

Peabody, M., Hutchens, N., Lewis, W., & Deffendall, M. (2011). First-Generation College Students at the University of Kentucky. Lexington, KY: Kentucky Education Policy & Law Lab. Available at: http://uknowledge.uky.edu/cgi/viewcontent.cgi?article=1000&context=packe

SERVICE

Invited reviewer, AERA Professional Certification & Licensure SIG: 2015
Invited reviewer, AERA Rasch SIG: 2015
Invited reviewer, Archives of Physical Medicine and Rehabilitation: 2013, 2014
Invited reviewer, Robinson Scholars Program Scholarship Selection Team, University of Kentucky: 2013, 2014
Invited reviewer, Mid-Western Educational Research Association, Measurement SIG Conference Proposals: 2013
Invited reviewer, Paul F. Fidler Research Grant, National Resource Center, University of South Carolina: 2013, 2014
University of Kentucky Scholarly Learning Community on First-Generation College Students: 2010-2011

PROFESSIONAL AFFILIATIONS

American Educational Research Association (AERA)
Division D – Measurement and Research Methodology
Division I – Education in the Professions
SIG – Professional Licensure and Certification
SIG – Rasch Measurement

SIG – Survey Research in Education
National Council on Measurement in Education (NCME)
Kentucky Association for Institutional Research (KAIR)

COMMUNITY

WGPL Neighborhood Association President: 2010, 2011
LFUCG Nicholasville Rd Steering Committee: 2009