Assessing Potential Predictors of Rater Fit Measures

in the Establishment of Performance Standards

By

MARIA INCROCCI

B.S., University of Illinois College of Pharmacy, Chicago, 1984
M.S., Chicago State University, Chicago, 2005

DISSERTATION

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2015

Chicago, Illinois

Defense Committee:

Carol Myford, Chair and Advisor
Lidia Dobria, Wright College
Everett Smith, Educational Psychology
Yue Yin, Educational Psychology
Dale Wurster, University of Iowa

ACKNOWLEDGEMENTS

I am deeply grateful to Dr. Carol Myford, my dissertation chair, for her support, time, expertise, and guidance throughout my academic experience and the dissertation process. Dr. Myford epitomizes excellence in education, and I have been privileged to be the recipient of her mentorship. I also acknowledge and thank the other members of my dissertation committee, Dr. Lidia Dobria, Dr. Everett Smith, Dr. Yue Yin, and Dr. Dale Wurster. Dr. Dobria was instrumental in validating the hierarchical model that I used for the data analysis, and I thank her for her time in evaluating the appropriateness of the model. I appreciate the guidance from Dr. Smith, Dr. Yin, and Dr. Wurster in the proposal stage of the research that resulted in formulating a solid foundation for the work. I am particularly grateful to Dr. Wurster (who had no idea when he raised his hand at a standard setting in 2011 expressing to the facilitator his level of (dis)comfort in rating items that were not necessarily in the content area of his expertise) for planting the seed for a future research project for a budding graduate student.

I am grateful to the National Association of Boards of Pharmacy (NABP), particularly Carmen A. Catizone, Executive Director, for the confidence and support provided to successfully complete a PhD program. Additionally, I am thankful for the permission to use data generated from an NABP-hosted standard setting.

And, finally, to my three amazing daughters, my friends, and my colleagues at NABP who have supported me throughout this journey, I thank you for your love, friendship, and commitment to me through a variety of challenging times.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF ABBREVIATIONS

BPS	Basic Biomedical and Pharmaceutical Sciences
CS	Clinical Sciences
FPGEC	Foreign Pharmacy Graduate Equivalency Certification
FPGEE	Foreign Pharmacy Graduate Equivalency Examination
HLM	Hierarchical Linear Regression Model
ICC	Item Characteristic Curve
IRT	Item Response Theory
MQC	Minimally Qualified Candidate
NABP	National Association of Boards of Pharmacy
PCOA	Pharmacy Curriculum Outcomes Assessment
SAS	Social/Behavioral/Administrative Pharmacy Sciences

SUMMARY

Licensing and certification organizations establish cut scores for examinations to specify the minimum levels of performance that candidates must demonstrate in order to be classified as competent. Therefore, a cut score should represent an appropriate performance standard for identifying proficient, knowledgeable individuals. Typically, a licensing or certification organization will convene a group of subject matter experts in the field (i.e., raters) and engage them in a standard-setting process to recommend a cut score.

Rater selection is a critical step in standard setting. Raters must be very familiar with the knowledge, skills, and abilities of the candidate population. To set a defensible cut score for an examination composed of multiple-choice items, raters must be able to judge accurately how minimally qualified candidates would likely perform. During a standard setting, it is common practice for raters to examine individual items and then provide an estimate of the proportion of minimally qualified candidates that the rater believes would answer each item correctly. Inevitably, there is variability among the raters in their judgments of candidate performance; some raters are able to provide more accurate proportion correct estimates than other raters. *Rater fit* refers to the level of accuracy or precision that an individual rater attains when providing these estimates. Using the raters' proportion correct estimates and calculated probabilities of a correct response for an item at the cut score, one can calculate *rater fit indices* that indicate how accurate each rater was when making a judgment about the performance of minimally qualified candidates on each item.

The purpose of this study was to determine to what extent two rater background-related variables (i.e., a rater's gender and content domain expertise) and two item characteristic-related

variables (i.e., an item's difficulty classification and content domain classification) could account for variance in rater fit indices. The fit indices were based on raters' proportion correct estimates of the performance of minimally qualified candidates on a 200-item certification examination that the National Association of Boards of Pharmacy developed. The 24 raters who participated in the 2011 standard setting were faculty members who had taught in U.S. colleges and schools of pharmacy for at least ten years.

A hierarchical linear model was used to conduct a two-level (items nested within raters) analysis. The level-1 model included the two item characteristic-related predictor variables while the level-2 model included the two rater background-related predictor variables. The outcome variable was the rater fit indices.

The two item characteristic-related variables accounted for 91% of the variance in the rater fit indices, suggesting that the ability to provide accurate proportion correct estimates for minimally qualified candidates was related to an item's difficulty level and content domain classification. By contrast, the rater background-related variables explained very little of the variance in the rater fit indices, after taking into account the variance in the indices that the content domain classifications and the item difficulty classifications explained. The ability to provide accurate proportion correct estimates for minimally qualified candidates was not related to a rater's gender or content domain expertise.

The study's findings support the standard-setting experts' view that rater training which includes multiple practice rounds, discussions, interactions, and feedback can be influential in decreasing the variance in raters' estimates of the proportion of minimally qualified candidates

who would answer an item correctly. Additionally, the study's findings reinforce the importance of providing plenty of opportunities during training for raters to make judgments about candidate performance on items from different content domains, as well as items that differ in their levels of difficulty.

1. REVIEW OF LITERATURE

1.1 Introduction/Background

Licensing and certification organizations establish performance standards to specify the minimum level of performance that candidates must demonstrate in order to be classified as competent. Meeting or exceeding a performance standard generally results in the consent to practice within one's profession by virtue of having obtained a license or certificate. Typically, a license confers the legal authorization to practice as determined by the licensing body. In many cases, an individual may opt for certification as an extension of licensure. In some disciplines, certification may denote that an individual has a specialization in a particular area or has demonstrated a level of knowledge necessary to partially fulfill the requirements of certification.

Test scores from high-stakes licensing and certification examinations must be interpretable as valid measures of candidates' knowledge, skills, and abilities to practice, given that the examination content is reflective of the job-related tasks or other essential areas (knowledge) critical to certification. It is important that a *cut score* (i.e., a score at a particular point along a continuum that defines minimally competent performance) represents an appropriate performance standard for identifying proficient, knowledgeable individuals. Typically, experts in the field recommend cut scores using a standard-setting process. The recommendations for performance standards resulting from standard-setting processes reflect expected levels of competence as determined by the judgment of experts in the field (Kane, 1994). Simply put, a standard setting represents the process that guides entities such as professional licensure agencies and educational boards in the establishment of cut scores. The

Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) make clear the importance of setting defensible cut scores:

> Cut scores embody value judgments as well as technical and empirical considerations. Where the results of the standard-setting process have highly significant consequences, those involved in the standard-setting process should be concerned that the process by which cut scores are determined be clearly documented and that it be defensible. (p. 101)

In addition, the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978) maintain that scores on examinations used for the purpose of selecting or identifying competent candidates must be "predictive" of essential functions that are deemed necessary in order to demonstrate competent performance of job-related tasks. When establishing a performance standard, a licensing or certification organization must provide evidence that it used a sound and reasonable standard-setting process.

Examinations serve as one of the criteria that regulators use to make decisions about an individual's eligibility for licensure or certification. Licensing and certification organizations routinely use examinations to assess the knowledge, skills, and abilities of candidates in relation to their prospective area of practice (Raymond & Neustel, 2006), and a candidate's scores on these examinations must accurately reflect his or her abilities to apply those knowledge and skills in practice (Clauser, Margolis, & Case, 2006). In addition, the societal implications of making inappropriate interpretations and uses of test scores are of paramount concern to licensing and certification organizations whose mission is to gain (and maintain) the public's trust (and end users such as regulatory bodies) in their respective professions (Haladyna, 1994).

When licensing and certification organizations assemble standard-setting panels of raters, they use systematic, purposeful sampling procedures to try to ensure that the panel represents the population of credentialed individuals in terms of their professional backgrounds and experiences. Guidelines from the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) suggest that "a sufficiently large and representative group of participants should be involved to provide reasonable assurance that the expert ratings across judges are sufficiently reliable and that the results of the judgments would not vary greatly if the process were replicated" (p. 101). The raters must conceptualize the characteristics of a *minimally qualified candidate* (MQC). During the standard-setting process, the raters engage in discussions in an effort to arrive at a common understanding of what a sufficiently knowledgeable, yet MQC knows and can do. Their definition provides a foundation or "common ground" that they use as they make judgments about candidate performance on individual test items. Most standard-setting approaches require that raters use the conceptualized image of the MQC to guide them as they render their judgments regarding candidates' performances on test items (Jaeger, 1995).

Best practices dictate that raters involved in standard-setting processes need to be representative of the population of practitioners or experts in the field. While licensing and certification organizations recognize the importance of the panelist-raters to the standard-setting process, assembling a representative standard-setting panel presents a challenge. These organizations select from a pool of practitioners/experts who have varying levels of expertise and years of experience within their field. The organizations must consider the purpose of the examination when deciding whom to invite to serve on a standard-setting panel. For example, for an entry-level credentialing examination, the organization must decide whether it would be more appropriate to invite recently credentialed practitioners to serve on the panel, or more experienced

practitioners. When selecting a panel, the organization must also strive to represent the diversity of practitioners within the field in terms of key demographic characteristics, such as age, gender, and race/ethnicity.

Unfortunately, there is a dearth of research to help licensing and certification organizations make informed choices as they constitute their standard-setting panels. Few researchers have studied how rater-related characteristics may influence the judgments that raters make when they are engaged in standard setting. Are there aspects of raters' backgrounds and experiences that seem to make a difference in how they perform this task? Are some raters better able to conceptualize a MQC and provide more accurate estimates of the proportions of MQCs who are likely to be able to answer correctly items on an examination? If so, which raters are more skilled in making such judgments, and what are their backgrounds and experiences?

Licensing and certification organizations attempt to standardize the experience for all those involved by providing raters with training and practice before they participate in the actual standard setting; however, little is known about the effectiveness of such training procedures. Understanding how rater and item attributes (if any) influence the experiences of raters, and thus the outcomes of a standard setting, could provide licensing and certification organizations with information vital to the success of the standard-setting process.

1.1.1 **Different Approaches to Standard Setting**

In *test*- or *item-centered* standard-setting approaches (Jaeger, 1989), raters use item content and statistical information to help them make their judgments of candidate performance. That is, they evaluate item performance with respect to a given population of candidates who have taken the examination. As raters are forming their judgments about the performance of the candidates on the items, they may be presented with empirical information regarding how the items functioned

on the examination (e.g., item difficulty values, item discrimination values, item characteristic curves). They may also be given feedback regarding the inter-rater consistency of their judgments. After they have proposed a tentative cut score for the examination, they may be shown what the impact of their decision would be on the candidates who took the examination (i.e., the probable pass/fail rate), and they can then use that information in a subsequent round of standard setting to refine their decision, if needed.

By contrast, in *person-centered* standard-setting approaches, raters render their judgments based on their expectations of how a defined group of candidates would perform on each item. That is, they evaluate candidate performance with respect to a particular set of items. Raters consider the performance of a group of candidates with whom they are familiar (or samples of their work) when setting a cut score on an examination. Educational testing organizations commonly use person-centered approaches in standard settings, while licensing and certification organizations commonly employ item-centered approaches (Hambleton & Pitoniak, 2006). The Angoff (1971) standard-setting approach, an item-centered approach, and modified versions of this approach are commonly used in licensing and certification testing (Meara, Hambleton, & Sireci, 2001; Plake, 1998). These approaches employ probabilities and proportion correct measures provided by raters on a standard-setting panel to estimate the performance of a MQC on a set of test items. Each rater evaluates each item presented, estimating the proportion of MQCs who, in the rater's opinion, would answer the item correctly.

1.1.2 **Influences on Rater Judgments in Standard-setting Processes**

The selection of the raters is a critical step in the standard-setting process. In licensing and certification, the raters should be experts in the subject matter represented in the examination. Incumbent licensed (certified) professionals and educators in the content areas provide the

necessary expertise to ensure the integrity of the process (Norcini, Shea, & Kanya, 1988), and using content experts as raters in setting performance standards is common practice (Plake, Impara, & Potenza, 1994). Their role in the standard setting is to use their familiarity with the knowledge, skills, and abilities of the candidate population to make a recommendation for a performance standard that represents an appropriate criterion for passing the test (Geisinger & McCormick, 2010).

Researchers have shown that there is variability in raters' judgments when estimating the probability that a MQC will answer an item correctly (Haertel, 2008; Jaeger, 1991). Contributing factors include variability in raters' understanding of the purpose and importance of the standard setting, their conceptualizations of performance-level descriptors, as well as their conceptualizations of what a MQC should know and be able to do (Skorupski & Hambleton, 2005). How raters perceive and comprehend the description of the MQC will affect the outcomes of a standard setting (Skorupski, 2012). As raters become more comfortable with the concept of a MQC through ongoing discussions, the likelihood of their reaching consensus in their interpretation of the meaning of minimal competence improves (Skorupski & Hambleton, 2005). Consensus among raters may result in less variability across the group in their probability estimates (Hurtz & Auerbach, 2003) and higher inter-rater reliability in their judgments.

Raters commonly receive training to help them describe the characteristics of a MQC. A recognized challenge that raters face when using Angoff standard-setting approaches is conceptualizing and reaching consensus on what the characteristics of a MQC are, prior to making any predictions regarding their performance on items (Plake, 1998). Additionally, raters often experience difficulty in accurately estimating the probability that a MQC would get an item correct (Brandon 2004; Impara & Plake, 1998). Critics of Angoff approaches question the

accuracy and consistency of raters' judgments regarding how MQCs *would* perform on items (Goodwin, 1999; Reid, 1991). It is likely that there will be variability among the raters in their judgments of candidate performance, the critics contend, thus introducing variance not accounted for in the measures utilized for setting performance standards (Norcini, 1994).

When using Angoff standard-setting approaches, raters often engage in informative processes that provide support for their initial proportion correct estimates (or, conversely, provide support for altering those estimates). These processes may include group discussions about the characteristics of the MQC, the presentation of empirical data revealing how a defined population of candidates performed on the items, as well as the presentation of information regarding the amount of agreement among the raters in their proportion correct estimates (Reckase, 2001). Providing item performance data to the raters affords them an opportunity to assess their accuracy in estimating the probability of a MQC getting the item correct by comparing their proportion correct estimates to actual empirical data. Facilitating discussions among raters results in stronger consensus in their judgments regarding the targeted population of candidates' performances on the items, which translates into less variability in raters' judgments (Hurtz & Auerbach 2003; Skorupski & Hambleton, 2005). Researchers have shown that providing raters with actual empirical data significantly improved the accuracy of raters' proportion correct judgments (Clauser et al., 2009), thus improving their estimations of proportion correct responses for a targeted population of candidates.

1.1.3 **The Potential Role of Rater Fit Indices in Helping to Build a Validity Argument**

In the context of this study, *rater fit* refers to the level of accuracy or precision that an individual rater attains when making proportion correct estimates of the performances of MQCs on test items. Licensing and certification organizations could benefit from the identification of

rater and item characteristics that influence a rater's ability to provide sound, reliable proportion correct estimates. They could use this information in several ways.

First, having knowledge of rater characteristics that influence standard-setting judgments could prove valuable when selecting raters to serve on a standard-setting panel. For example, suppose that an organization carried out a standard setting for a credentialing exam for entry-level practitioners. When they analyzed the raters' proportion correct estimates, they found that raters who had been working in a field for many years tended to be less accurate in their proportion correct estimates (i.e., had poorer rater fit) than raters who had fewer years of experience in that field. The organization could use that knowledge in the future to help them select practitioners to serve on standard-setting panels.

Second, the organization could use knowledge of rater characteristics that influence standard-setting judgments to revise and strengthen their rater training program. If they choose to include on the standard-setting panel some practitioners who have been working in the field for a number of years, they might decide to allocate more training time to helping to ensure that all the practitioners reach a common understanding of what a minimally qualified entry-level practitioner should be expected to know and be able to do. The goal would be to make certain that all practitioners on that panel--those with few years of experience in their field, as well as those with many years of experience--arrive at a common conceptualization of the entry-level practitioner before they begin the standard-setting process.

Performance standards are not "created" during a standard setting. A series of well-planned activities leads to expert raters making informed judgments, resulting in the establishment of a recommendation for setting a cut score. Policy makers then evaluate the outcomes of a standard setting and the cut-score recommendation of the standard-setting panel. It is the policy makers

who are ultimately responsible for rendering a final decision regarding what the cut score will be. Consequently, the setting of a cut score is a policy decision informed by a lengthy and complex process of defining the test's purpose, assembling a panel of expert raters to carry out a defensible standard setting, and then appropriately interpreting the outcomes of that process.

Licensing and certification organizations that carry out standard-setting processes need to provide evidence to show that they conducted those processes in a reasonable, systematic, and thoughtful manner to arrive at a defensible cut score. As Kane (2001) explained, the documentation of the processes and tasks executed during a standard setting provides critical sources of validity evidence that a licensing or certification organization can employ to support the use of test scores for their intended purpose. The procedures used to select the standard-setting raters, the rationale for the selection and training of raters who participate in a particular standard-setting approach, the description of how empirical item information is introduced, the content of the discussions in which the raters engage, the feedback that the raters provide regarding their levels of satisfaction with the standard-setting process and how it was conducted are all critical, documentable details of the standard-setting process that can contribute to the validity argument, supporting the interpretations of the outcomes and recommendations.

If licensing and certification organizations had access to fit indices for the raters participating in their standard-setting panels, they could use that information to help in building a validity argument to support some of their decision-making processes for the selection and training of panelists. It is incumbent on licensing and certification organizations to gather validity evidence to support decisions that will directly influence the interpretation of test scores. Using systematic and well-documented processes that support the methodological and theoretical

foundations of the standard-setting process would provide licensing and certification organizations with defensible evidence that they are adhering to best practices.

Policy makers rely on measurement experts to ensure that the standard-setting processes uphold the standards set forth by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The consequences of using subjectively derived or unjustifiable cut scores would undermine the integrity of the examination program. In licensing and certification, the regulation of candidates for the purpose of protecting the public would be vulnerable to public scrutiny.

1.1.4 **Rater-related Variables that May Affect Measures of Rater Fit**

There may be rater-related variables that affect measures of rater fit. For example, raters having different areas of specialization may vary in the levels of comfort they experience when performing the judgmental tasks required during a standard setting. There may also be differences in the ways in which raters from various areas of specialization approach the standard-setting task, and in the thought processes they use to arrive at their judgments. Perhaps raters' prior experiences in their fields may influence their abilities to accurately estimate the proportion correct responses of MQCs. If researchers were to find that one or more of these variables influence raters' judgments, then it would be important to consider those variables when identifying practitioners to serve as raters in a standard-setting panel. Somewhat surprisingly, few researchers have investigated these types of rater-related variables and the impact that they might have on the outcomes of a standard setting.

Researchers working with the American Board of Internal Medicine conducted a small study involving medical doctors who were selected to serve as raters in a standard setting for one of the board's examinations. The researchers reported that there were no differences in the

doctors' judgments of the performance of a conceptualized MQC, despite the fact that the doctors differed in their areas of specialization (Norcini, Shea, & Kanya, 1988). The researchers speculated that the discussions among the raters that addressed characteristics of the targeted candidate population provided the necessary "neutralizing" effect that brought the raters to a common conceptualization of a MQC, despite the differences in the experts' backgrounds. Clauser et al. (2009) studied how practicing physicians participating in a standard setting used information they were given about the performance of the candidates on an examination (i.e., the probabilities of correctly answering an item for candidates in five score groups, and the proportion of candidates in each of the five groups who chose each of the five options for each item). The researchers also examined how the raters' interactions impacted their judgments of candidate performance. Group discussions reduced the variability in the raters' estimations of candidates' probabilities of success but did not improve the accuracy of their estimates of the relative difficulty of the test items. However, providing the raters with information about how the candidates performed on the items significantly improved the precision of their judgments of candidates' probabilities of success when compared to the empirical probabilities. The researchers concluded that if the goal of a standard setting is to strive for rater agreement in their judgments, then providing raters with information about how the candidates performed on the items can assist raters in that task.

The few studies that have focused on gaining an understanding of the cognitive processes that raters employ during standard settings have shown that interaction, discussion, and feedback assist in building continuity among raters, reducing inter-rater variability in their judgments (Brandon, 2004). When provided with opportunities to change their proportion correct estimates,

raters have indicated that discussions with the other raters were more influential than reviewing information about how candidates performed on items (Noricini, Shea, & Kanya, 1988).

1.1.5 **Item-related Variables that May Affect Measures of Rater Fit**

When raters are judging how MQCs will perform on test items, some of the characteristics of those items may influence the raters' judgments. Item characteristics that might affect measures of rater fit could include one or more of the following: (a) the difficulty of the item, typically defined as the proportion of candidates who got the item correct (*p*-value); (b) the cognitive level of the item, typically identified using a taxonomy such as Bloom's Taxonomy of the Cognitive Domain; and (c) the specificity or level of detail of the content addressed in the item. During Angoff-based standard settings, providing raters with some information regarding the characteristics of the items (e.g., *p*-values) gives them a realistic sense of how the population of candidates taking the examination performed. Comparing raters' probability estimates to actual empirical data provides a means for evaluating the accuracy of the raters' judgments.

Raters differ in the levels of precision they can attain when they are estimating the proportion of MQCs who would be able to answer an item correctly (Impara & Plake, 1997; Reid, 1991; Taube, 1997), casting some doubt on the ability to replicate any given standard-setting process. Some researchers have reported that the raters they studied tended to overestimate the proportion of MQCs who would answer a difficult item correctly, while they underestimated the proportion of MQCs who would answer an easy item correctly (Shepard, 1994; Taube, 1997). However, other researchers (Impara & Plake, 1998) have failed to replicate those findings, arguing that there is no significant relationship between raters' proportion correct estimates and the difficulties of items.

One option to try to reduce inaccuracies in raters' estimations of MQC performance is to provide raters with item *p*-values, or similar measures of difficulty, during the standard setting. Pooling data from the full distribution of candidates (an overall *p*-value) provides information that is generalizable but does not necessarily align with how the MQC would perform. Raters may be able to use the candidate performance data as information to guide them in making their proportion correct estimates. Intuitively, it would seem reasonable to provide *p*-values from a subset of candidates whose performances was close to the cut score. However, the process used to identify the range of test scores that appropriately define "near the cut score" may be subjective. In addition, there may only be a small number of candidates "classified" as MQCs, thus minimizing the number of candidate scores to use reliably as anchors for item-level data (Clauser et al., 2009). Nonetheless, some researchers contend that raters' proportion correct estimates will be more accurate if raters are made aware of the difficulties of items than if they make their judgments in the absence of such data (Norcini, Shea, & Kanya, 1988). Not all researchers see the merit in such a strategy, though. For example, Taube (1997) cautioned that having raters rely on empirical data to make their judgments instead of their conceptualizations of what a MQC should know and be able to do may taint the fundamental purpose of using expert raters to identify a performance standard.

1.2 **Purpose of the Study**

The purpose of this study was to determine to what extent two rater background-related and two item characteristic-related variables could account for the variance in rater fit indices. Raters in my study participated in a standard setting in which they made judgments about candidate performance on an examination used in a certification process. The rater fit indices that I

used were based on the raters' proportion correct estimates of the performance of MQCs on the test items.

I used standard setting and item performance data from a certification examination that the National Association of Boards of Pharmacy (NABP) developed. The purpose of the standard setting was to establish a defensible performance standard for an examination designed to test an internationally trained pharmacist's knowledge of a United States Pharmacy curriculum. The Foreign Pharmacy Graduate Equivalency Examination (FPGEE) is one of the criteria used in the determination of the eligibility to obtain a Foreign Pharmacy Graduate Equivalency Certification (FPGEC). The United States boards of pharmacy recognize obtaining this credential as a required step in the process of becoming a licensed pharmacist. Candidates who obtain the FPGEC certificate are also required to serve practice hours as licensed interns and subsequently must pass the North American Pharmacist Licensure Examination and the appropriate jurisprudence examination in the state in which they are seeking primary licensure.

It has become customary in licensure and certification testing to use an item response theory (IRT) approach (Hambleton & Swaminathan, 1985; Lord, 1980; Wright & Stone, 1979) for the assembly, analysis, scoring, and reporting of results from tests. When using an IRT approach, the probability of a candidate's correct response to an item is a function of the candidate's competency (referred to as a latent trait, θ) and one or more characteristics of the item (i.e., difficulty, discrimination, guessing). While organizations use IRT approaches for examination assembly, analysis and scoring, they tend to use non-IRT approaches for the establishment of performance standards (Ferdous & Plake, 2008). In my study, I used data from a certification examination program that employs a traditional (non-IRT) approach to establish performance standards and an IRT approach to creating the test and reporting candidate scores.

1.3 **Primary and Secondary Research Questions**

The primary research question that I sought to answer was this: To what extent can two rater background-related variables (i.e., gender, content domain expertise) and two item characteristic-related variables (i.e., item difficulty, content domain/subject area) account for the variance in rater fit indices?

Research questions that were of secondary interest were as follows:

1. Is there justification for making adjustments to a rater's proportion correct estimates (or eliminating a rater's estimates) if a rater's fit index calculated from those estimates exceeds a certain upper limit? If so, what should that upper limit be? How would adjustments be made?
2. In practice, how could knowledge of rater background-related variables and/or item characteristic-related variables that can explain the variance in rater fit indices inform decisions regarding the selection of experts to serve on standard-setting panels, training models used in standard settings, and standard-setting outcomes? In other words, if a licensing or certification organization could establish that experts having certain background characteristics had rater fit indices that exceeded some predefined limit, should the organization consider that information when deciding whom to invite to serve on a standard-setting panel? If a licensing and certification organization could establish that one or more item characteristic-related variables explains much of the variance in the raters' fit indices, how could the organization use that information to inform the design of rater training?

3. Could the gathering and reporting of rater fit-related evidence contribute to the development of a validity argument to support the interpretations of test scores for a specified use? If so, how?

1.4 Significance of the Study

A compelling challenge in using Angoff and modified Angoff standard-setting approaches is for raters to accurately estimate how MQCs will perform on test items (Impara & Plake, 1997). During the training of raters, MQCs are the focal group for raters to conceptualize. Raters engage in group discussion to try to arrive at a common understanding of what MQCs know and do not know. Studying the levels of accuracy that raters are able to attain when making their proportion correct estimates would help licensing and certification organizations become cognizant of how raters' judgments differ when they carry out this critical task (Goodwin, 1999; Smith & Smith, 1988).

Assessing rater fit and variables that may be predictive of fit (or misfit) could provide licensing and certification organizations with useful information that they could employ when deciding whom to include in standard-setting panels and in training raters to participate in standard setting. As an example, the FPGEE is comprised of four main content domains that represent the breadth of pharmacy education. In the 2011 FPGEE standard setting, the raters provided proportion correct estimates for MQCs for all items in all content domains. It may be that some raters' proportion correct estimates for items in their own content domain of expertise demonstrate better fit than their proportion correct estimates for items that are outside their content domain of expertise. If this were the case, then NABP might want to consider instituting a procedure to adjust raters' proportion correct estimates in those cases in which raters' fit indices are outside some predefined limit. Alternatively, NABP might consider using a different

approach for standard setting that isn't dependent on raters providing proportion correct estimates for MCQs.

Methods of establishing performance standards are under constant review for defensibility and adherence to best practices in educational measurement. It is the responsibility of licensing and certification organizations to provide validity evidence to support their claims that they carried out their standard settings in an appropriate manner (AERA, NCME, & APA, 2014). They must ask the hard questions regarding the "goodness of fit" for the standard-setting approaches utilized in their programs. In this study, I investigated the predictive properties of two rater background-related variables and two item characteristic-related variables, seeking to determine to what extent they might provide valuable information to help build a validity argument to support the use of scores on the FPGEE as one of the criteria for making decisions regarding candidates' eligibility to obtain Foreign Pharmacy Graduate Equivalency Certification.

1.5 **Review of Models to Estimate Rater Fit**

In modified Angoff standard-setting procedures, raters engage in a review of test items and provide proportion correct estimates for each item, which represent the probability of a MQC getting the item correct. An item characteristic curve (ICC) plots the probability of a correct response as a function of a candidate's level of competency θ. The ICC is a monotonically increasing function that describes the probability of correctly responding to an item as a function of the measured difficulty of that item. Summing and averaging raters' proportion correct estimates and mapping to an ICC provides a process for determining a performance standard for a given test (Plake & Kane, 1991).

An important consideration when using conversion methods such as mapping the proportion correct estimates to the latent trait scale is the soundness or accuracy of the probability estimates. Having access to measures of item difficulty provides a means for comparing raters in terms of their accuracy in estimating the proportion of MQCs who would be likely to answer each item correctly. Correlations of raters' proportion correct estimates to item difficulty measures that have been calculated from the responses of a group of MQCs to a set of items provide information to evaluate rater fit (Goodwin, 1999). The resulting IRT measures are on an equal-interval, linear scale. Each measure describes the relationship between a given candidate's competency measure (θ) and the raters' judgments of the probability of candidates at that level of competency being able to get an item correct. The height of the ICC corresponds to the probability of a correct response to the item for candidates at any given θ. By comparing rater proportion correct estimates for candidates at various points along the ICC, one can evaluate the alignment between known item difficulty measures and a rater's proportion correct estimates (Hurtz, Jones, & Jones, 2008).

1.6 Review of Rater Fit Models that Involve the Conversion of Raters' Proportion Correct Estimates to Measures on an IRT Scale

1.6.1 van der Linden's Error of Specification and Consistency Indices

To measure rater fit, van der Linden (1982) compared raters' estimated *p*-values (i.e., their proportion correct estimates) for items to information provided in the ICC. Each rater is associated with an error of specification (E_{Ir}). The error of specification represents the departure of a rater's proportion correct estimate for an item (i.e., *p* value) from the probability of a correct response to that item at a targeted θ on an ICC. The error of specification presents as

$$E_{Ir} = \frac{\sum_{i=1}^{n} |Mir - Pi(\theta*)|}{n} \quad (1)$$

where M_{ir} is the rater's proportion correct estimate for item *i* for candidates who are at a particular competency level (e.g., MQCs), and $P_i(\theta*)$ is the height of the ICC at the point on the competency scale θ* for item *i*. The index represents the average deviation from the IRT probability (based upon θ) of getting the item correct for *n* items across a test for a single rater. The mean of all errors of specification across a group of raters carrying out the same standard-setting task is denoted as E_{IR}. Intuitively, a low E_{IR} value would indicate less variance in the raters' proportion correct estimates than a high E_{IR} value.

The consistency index (C_{Ir}) is a measure of the maximum deviation of a proportion correct estimate for a given item as a function of a candidate's competency location on the ICC. For example, if the proportion correct estimate is .40 at a given θ on an ICC, then the maximum deviation of a proportion correct estimate could not exceed .40 (i.e., that would be a situation in which the rater judged that the MQC would have zero probability of answering the item correctly). The C_{Ir} is represented as

$$C_{Ir} = \frac{E_{I,max} - E_{Ir}}{E_{I,max}} \quad (2)$$

where

$$E_{I,max} = \frac{\sum_{i=1}^{n} E_{i\,max}}{n'} \quad (3)$$

$E_{I,\,max}$ is the maximum possible error of specification across a set of items, and is equal to the sum of the errors of specification for the individual items divided by the total number of items.

As with errors of specification, aggregating the indices of consistency across raters provides an overall index (C_{IR}) of consistency across a set of items. Higher C_{IR} values indicate

consistency across raters in terms of the degree of alignment of their proportion correct estimates with known item information (i.e., the raters differed comparatively little in their estimates of the proportions of MQCs who would provide a correct response to the item).

1.6.2 **Kane's Chi-square Statistic for Measuring Rater Fit**

Kane (1987) suggested using a chi-square statistic to measure overall rater fit. In this equation, M_{iR} is the mean of the raters' *p*-values (i.e., proportion correct estimates) for a given item and represents the height of the item's ICC at a defined θ. Kane proposed that when raters' proportion correct estimates for the MQC align with known probabilities of correct responses on items, then the M_{iR} represents good rater fit.

According to Hurtz and Jones (2009), the problems with using a chi-square statistic as a measure of rater fit are two-fold. First, the hypothesis tested (i.e., the raters' proportion correct estimates correspond perfectly to the IRT probability measures at given θs) is idealistic and improbable, as one would not expect empirical data to exhibit perfect model fit. Secondly, as sample size increases (i.e., increases in the degrees of freedom), the examination's statistical power increases so that if the raters' proportion correct estimates deviated even slightly from the IRT probability measures at given θs, one would likely conclude that the data did not fit the model. Having raised these concerns, Hurtz and Jones cautioned that "the statistical significance of the chi-square statistics [is not] a viable criterion for rater fit" (p. 125).

An improvement to Kane's M_{iR} statistic gives more weight to those items that had higher levels of rater agreement in their proportion correct estimates (Hurtz & Jones, 2009). Therefore, items with smaller variances in their proportion correct estimates would have greater influence in determining a performance standard. Theoretically, this model would triage items for their

usefulness in determining a performance standard by "hand-picking" and using those items that demonstrate the best rater fit.

There have been few published studies of rater fit in which the researchers' focus was on examining the alignment between rater-determined performance standards and IRT item/test information. Hurtz and Jones (2009) conducted a study comparing the utility of the van der Linden and Kane rater fit statistics. It is the only study in the literature (to date) that has investigated the alignment of rater proportion correct estimates to IRT *a*-, *b*-, and *c*- parameters for ICCs.

The data set that the researchers used was from the administration of an undergraduate-level, multiple-choice examination in the allied health field. The nine raters who participated in the standard setting were faculty members at colleges and universities who were very familiar with the subject matter tested.

Rater training involved providing an explanation of the purpose of the examination and then having the raters discuss how four defined student groups would likely perform (i.e., highly competent, competent, marginally competent, and weak). The raters were to consider the marginally competent group as the most critical for standard setting (i.e., this group represented minimal competence for pass/fail decisions). Raters estimated the proportion of students in each of the four groups who would likely answer each item correctly. The raters could nominate questions for discussion with their peer raters as they saw fit, and they were permitted to change their proportion correct estimates after the discussion, if they requested to.

The researchers evaluated a number of rater fit indices produced for each of the four groups, including Kane's chi-square index and van der Linden's E_{IR} and C_{IR} indices. When they used Kane's approach to assess the correspondence of the raters' proportion correct estimates to

the ICCs for the four competence thresholds, they found that all four chi-square indices were statistically significant, which the researchers interpreted as evidence of poor data/model fit (though they acknowledged that the chi-square statistic was "an insufficient criterion" (p. 138) for making that judgment). The researchers reported that the raters' proportion correct estimates for each of the four competence thresholds were markedly restricted when compared to the ICCs at those threshold values.

When the researchers examined the rater fit indices they produced using van der Linden's approach, they found that the E_{IR} was highest for the marginally competent group (0.189) and lowest for the highly competent group (0.122). This finding implies that the raters' proportion correct estimates demonstrated better fit to the ICCs when the raters were making judgments regarding the performance of more competent students than when they were making judgments regarding the performance of the other student groups. The C_{IR} was highest for the highly competent group (0.856) and lowest for the weak group (0.723), with the marginally competent group (0.730) running a very close second. That is, the raters were more consistent in their proportion correct estimates across items when judging the performance of highly competent students than when judging the performance of less competent students. Raters tended to overestimate the performance of minimally competent students, a finding that other researchers have reported when they conducted similar studies in which they investigated the relationships between raters' proportion correct estimates and empirical item information (i.e., *p*-values)(Bejar, 1983; Goodwin, 1999; Impara & Plake, 1998).

After comparing a number of possible indices of rater fit, Hurtz and Jones (2009) concluded that van der Linden's E_{IR} and C_{IR} indices "appear to be superior to methods based in Kane's writing, which involve the standard errors of the ratings" and to those indices that were

"based in the writings of Impara and Plake (1996) and Goodwin (1999), which involve defining fixed intervals around the ICC to define accurate ratings" (p. 141). While there have only been a few studies in which researchers have investigated rater fit using proportion correct estimates and known item parameters, the findings from those studies would seem to suggest that van der Linden's error of specification and consistency indices appear to be viable options to investigate further.

Clearly, there is a need for more research to consider the utility of these indices for understanding rater fit and the role that it may play in helping to design and implement more effective standard-setting approaches. I designed my study to help to address this gap in the rater fit literature. My plan was to investigate several variables that may be related to the level of accuracy a rater attains when providing estimates of the proportion of a targeted candidate population that would answer a test item correctly. Specifically, the principle aim of my study was to determine to what extent two rater background-related variables (i.e., gender, content domain expertise) and two item characteristic-related variables (i.e., item difficulty, content domain/subject area) accounted for variance in rater fit indices.

2. METHOD

I used a hierarchical linear regression model (HLM) to determine the extent to which two rater background-related variables and two item characteristic-related variables could explain the variance in raters' fit indices. Hierarchical linear modeling relaxes distributional assumptions of normality in data (particularly when considering the outcome variable), making it an appropriate approach for investigating the relationships among my set of variables. Additionally, when data have a hierarchical nature (e.g., items nested within raters), one can account for the complete structure of the data set in the analysis.

2.1 **Participants**

In 2011, NABP convened a panel of subject matter experts (i.e., "raters") to participate in a standard-setting session for the FPGEE. NABP solicited the raters from a database of U.S. college/school of pharmacy academicians. These individuals had some affiliation with the FPGEE program (i.e., each had at least some past experience as a subject matter expert during item development and/or had been appointed to, and held tenure on, the program review committee). There was a concerted effort to identify individuals who had taught in a pharmacy program for five years or more in at least one of the four content domains represented in the examination. NABP sought to convene a panel of raters who represented academic institutions that were located in various parts of the country. The solicitation letter defined the role of the rater and made it clear that NABP expected all prospective raters to participate in a training session and complete all of the exercises encountered during the standard-setting meeting.

Additionally, NABP expected each prospective rater to have a comprehensive understanding of the knowledge/competency statements upon which the FPGEE was based.

The panel was composed of 15 former item writers and nine FPGEE review committee members. The raters completed a demographic information form in which they provided their name, year of licensure (applicable to pharmacist licensure), educational credentials, workplace, gender, race/ethnicity and the state in which they resided/taught. Twenty-one of the raters held an entry-level degree in pharmacy, and all of the raters held advanced graduate degrees at the masters or doctorate level.

Each rater had been a faculty member in one (or more) U.S. pharmacy programs for a minimum of ten years. The primary teaching responsibilities and areas of expertise for seven of the raters were in Content Domains 1 and 2 (i.e., Basic Biomedical and Pharmaceutical Sciences--BPS). Five of the raters declared expertise in Content Domain 3 (i.e., Social/Behavioral/Administrative Pharmacy Sciences--SAS), and the remaining 12 raters declared their area of expertise in Content Domain 4 (i.e., Clinical Sciences--CS).

2.2 **Materials**

The FPGEE knowledge (competency) statements are reflective of the U.S. College of Pharmacy Curriculum. (See Appendix A.) NABP conducts periodic surveys of the colleges of pharmacy in order to monitor trends in pharmacy programs. The most recent survey conducted in 2010 resulted in a 65% response rate, providing information about a variety of pharmacy programs. The survey included questions about the topics covered in courses taught in the pharmacy curriculum and the weight or influence of individual courses relative to the total curriculum. NABP collected data on the number of credit and semester hours appropriated to

each of the topics within courses and weighted them to produce a hierarchical representation of the content domains with regard to the time spent in the courses. A committee of college of pharmacy academicians reviewed the survey outcomes and made recommendations to the NABP Advisory Committee on Examinations and the NABP Executive Committee regarding the distribution of content for the FPGEE examination.

The operational FPGEE is a 200-item test with 50 non-scored field-tests items. All items are single-answer, multiple-choice items, with each item having four options. The FPGEE covers four main content domains (i.e., Basic Biomedical Sciences, Pharmaceutical Sciences, Social/Behavioral/Administrative Pharmacy Sciences, and Clinical Sciences). The examination is available two times a year, typically in the spring and fall. The examination is only available in the U.S. and administered through vendor-contracted testing centers. Candidates who live outside of the country are required to travel to the U.S. to take the examination. The time allotted for the examination is four-and-one-half hours. Candidates take the examination in two sessions with a mandatory break separating the sessions.

To facilitate the standard setting, NABP assembled a reference form containing 200 operational items. The process used to assemble the reference form for the standard setting mirrored forms assembly for the operational FPGEE. The reference form met all psychometric targets (i.e., test characteristics curve, test information function) and content (i.e., exam blueprint) constraints.

2.3 Classifying the Items and the Raters (i.e., Creating the Predictor Variables)

The predictor variables for this study included two rater background-related variables (i.e., the rater's content domain expertise and gender) and two item characteristic-related variables (i.e., the item's difficulty and the content domain of the item).

For the first rater background-related predictor variable, I created three classifications to characterize the raters' content domain expertise. The three content domain expertise codes were Basic Biomedical and Pharmaceutical Sciences (BPS), Social/Behavioral/Administrative Sciences (SAS), and Clinical Sciences (CS). When creating the classifications, I combined Basic Biomedical and Pharmaceutical Sciences. The subject matter experts in these two content domains routinely work together as a single group during test development processes to author, review, and code items. Additionally, as a group, they conduct reviews of Basic Biomedical and Pharmaceutical Sciences items that appear on the examination forms. Given how these subject matter experts function, it seemed reasonable to combine those two content domains into one classification. The nine raters who are members of the FPGEE review committee had previously declared their respective areas of content domain expertise. (They routinely review test items within those domains.) I asked the remaining 15 raters to declare their areas of content domain expertise. All but one of the raters complied with my request. In order to assign that rater an appropriate content domain expertise code, I reviewed his curricula vitae and faculty roster pages posted on the American Association of Colleges of Pharmacy website to confirm his primary teaching responsibilities in the U.S. pharmacy curriculum. The raters' content domain expertise classifications were as follows: Basic Pharmaceutical Sciences (n = 6), Social/Behavioral/Administrative Sciences (n = 6), and Clinical Sciences (n = 12).

The second rater background-related predictor variable was gender. When the raters completed the demographic information form, they indicated their gender. The raters' gender classifications were as follows: female (n = 10) and male (n = 14).

For the first item characteristic-related predictor variable, I classified each item according to its content domain. Subject matter experts (authors) had previously coded the items and mapped them to the FPGEE competencies, and the program review committee had verified the item coding. I classified the items using the same three codes that I used for the raters: Basic Biomedical and Pharmaceutical Sciences (BPS), Social/Behavioral/Administrative Sciences (SAS), and Clinical Sciences (CS). Again, because it is customary for subject matter experts in the Basic Biomedical Sciences and Pharmaceutical Sciences to work together to develop and review items, I combined these two content domains to create a single classification. The item content domain classifications were as follows: Biomedical and Pharmaceutical Sciences (93 items), Social/Behavioral/Administrative Sciences (45 items), and Clinical Sciences (62 items).

For the second item characteristic-related predictor variable, I used p-values (i.e., measures of item difficulty based on the proportion of students who correctly answered each of the items) to classify the items. To obtain these measures, I analyzed the responses of a set of "borderline" students to those items. They were all U.S. College of pharmacy students who were either in the third or fourth year of the professional curriculum. (The four years of professional curriculum follow at minimum two years of preparatory coursework.) NABP collected this data in 2011 during administrations of a Pharmacy Curriculum Outcomes Assessment (PCOA), which is only available to U.S. colleges of pharmacy students. The competency statements that defined the test blueprint for the PCOA were the same as the competency statements that defined the test

blueprint for the FPGEE. To assemble the FPGEE standard-setting form, NABP used items that they had previously tested on U.S. students.

In preparation for the FPGEE standard setting, NABP employed three sets of item *p*-values for three cohorts of students: (a) all students, (b) high scoring students (relative to the FPGEE cut score), and (c) borderline students who scores were near (i.e., at or slightly above) the FPGEE cut scores. For this study, I used the set of 200 item *p*-values that were based upon the performance of "borderline" students whose scores were at, or slightly above, the FPGEE cut score (θ*). On average, the responses of 100 borderline students informed the calculation of the *p*-value for each of the items.

Next, I rank ordered the 200 item *p*-values to determine the spread of the difficulty measures. The item *p*-values ranged from 0.077 to 1.00 (M = 0.521, SD = 0.211). I then classified the items into four groups based on the departure of each item's *p*-value from the mean item *p*-value. Conveniently, I was able to classify all item *p*-values into four groups. Twenty-five item *p*-values were between -2 SD and -1 SD from the mean *p*-value (i.e., *difficult* items), 75 item *p*-values were ≥ -1 SD from the mean *p*-value but < 0.521 (the mean *p*-value) (i.e., *moderately difficult* items), 73 item *p*-values were ≥ 0.521(the mean *p*-value) but < 1 SD from the mean *p*-value (i.e., *moderately easy* items), and 27 item *p*-values were ≥ 1 SD from the mean *p*-value but < 2 SD from the mean *p*-value (i.e., *easy* items).

My rationale for classifying the items by their *p*-values was to differentiate among them in terms of their relative difficulty. I treated item difficulty as a categorical variable rather than as a continuous variable in the hierarchical linear model (HLM) analysis so that I would be able to investigate the relationships between the four *p*-value classifications of the items (i.e., a predictor variable) and the variance in the rater's proportion correct estimates (i.e., the outcome variable).

Treating item difficulty as a categorical variable may help with the interpretation of outcomes and may be justified when there is question of whether the relationship between the predictor variable (*p*-values) and the outcome variable (rater fit indices) is linear (DeCoster, Gallucci, & Iselin, 2011). When assembling FPGEE test forms, NABP routinely uses item *p*-values to inform decisions regarding which items to include. They try to ensure that each test form contains a balance of difficult, moderately difficult, moderately easy, and easy items. Using the four classifications that I created resulted in a normal distribution of items that had *p*-values between -2 and + 2 SD from the mean *p*-value.

2.4 **Rater Training**

The raters participated in a three-hour training session to prepare for the standard setting. The training began with an overview of test development and the purpose of standard setting. The raters engaged in discussion to help them reach a common understanding of what minimally qualified, yet sufficiently knowledgeable, U.S. pharmacy students who had completed the didactic portion of their education would know.

Using an Angoff standard-setting approach, the raters participated in several practice rounds, providing estimates of the proportion of minimally qualified U.S. pharmacy students who would answer each item correctly. (The items used for training were not included in the actual reference form used in the standard setting.) The raters had access to three sets of item *p*-values for each of the practice items (i.e., for each item, a *p*-value for all students, a *p*-value for borderline students, and a *p*-value for high scoring students).

After completing the training, the raters read and signed an attestation that they understood their role in the standard-setting process, and that their training was adequate to prepare them for the standard setting.

2.5 **Data Collection Procedure**

Each rater provided proportion correct estimates for 200 operational items during two rounds. A proportion correct estimate represented a rater's appraisal of the proportion of minimally qualified, yet sufficiently knowledgeable, U.S. pharmacy students who had completed the didactic portion of their education who would answer the item correctly. All raters judged all items, regardless of a rater's content domain expertise or an item's content domain classification.

After the first round, the raters discussed what the new performance standard would be if based solely on the proportion correct estimates that the raters provided. They reviewed a subset of the items (i.e., those for which there was a great deal of variability in their proportion correct estimates) and shared their rationales for their judgments. The raters reviewed the *p*-values that reflected the performance of the borderline students on that subset of items. The discussion surrounding the borderline students' performance provided an opportunity for the raters to consider the opinions of their colleagues who were more familiar with the content of each of those items, given their particular content domain expertise.

Following this discussion, the raters then provided their second round of proportion correct estimates for the items in the standard-setting form. NABP collected, recorded and analyzed the data from this round and again provided the group with impact data. The impact data included the number of items a candidate would need to answer correctly in order to pass

the examination. In addition, they examined the impact of applying their performance standard on a cohort of FPGEE candidates, noting what the pass rate would be.

2.6 Data Analysis Procedures

2.6.1 Calculating Rater Fit Indices

To fix item and rater notation, let $i = 1, \ldots, I$ index the items and $r = 1, \ldots, R$ index the raters. Under the dichotomous Rasch model (Rasch 1960/1980), the probability $P_i(\theta^*)$ of a candidate answering an item correctly is a function of that candidate's level of competency (θ^*) and the item's level of difficulty (D_i). The equation is represented as:

$$P_i(\theta^*) = \frac{e^{(\theta^* - D_i)}}{1 + e^{(\theta^* - D_i)}} \quad (4)$$

where $P_i(\theta^*)$ is the height of the item characteristic curve (ICC) for item *i* at the cut score (i.e., the probability that a MQC answers item *i* correctly), (θ^*) represents the location on the competency continuum that denotes passing the examination (i.e., the competency level that a candidate must meet or exceed in order to be considered minimally qualified), and D_i represents the difficulty of a given item *i.*

Using the cut score (θ^*) calculated from the raters' actual 2011 FPGEE standard-setting proportion correct estimates and calibrated item difficulty measures (i.e., Rasch item calibrations), I calculated the probability of getting each of the 200 items correct for a MQC who had a competency measure at the cut score.

Next, I calculated van der Linden's (1982) *error of specification* (E_{Ir}) (modified by Hurtz and Jones, 2009) for each rater using the following equation:

$$E_{Ir} = \frac{\sum_{i=1}^{I} |M_{ir} - P_i(\theta^*)|}{I} \quad (5)$$

where M_{ir} is rater *r's* proportion correct estimate for item *i* for a MQC of competency θ^*, and $P_i(\theta^*)$ is as previously defined (i.e., the probability that a MQC answers item *i* correctly). Equation (5) sums the absolute values of the difference between M_{ir} and $P_i(\theta^*)$ for each item and then divides that total by 200 (i.e., the number of items on the standard-setting form). Thus, for a given rater, the *error of specification* (E_{Ir}) represents the variance in that rater's proportion correct estimates *over all the items* on the standard-setting form--a measure of *rater fit*. This is a summary measure of how accurately a rater was able to judge the performance of MQCs on the 200 FPGEE items. Intuitively, the lower the error of specification, the more accurate the rater's proportion correct estimates. Table I (Appendix B) shows each rater's error of specification (E_{Ir}), as well as the rater's classification according to the background-related variables of interest.

The error of specification index is averaged across all items, and thus it provides a global measure of each rater's fit across all items *I*. To investigate rater fit at the individual item level, I employed the numerator of the error specification index as the outcome variable in the analyses that follow. This version of the rater fit index has the following form:

$$VARIANCE_{ir} = |M_{ir} - P_i(\theta^*)| \quad (6)$$

For a given rater, *r*, this index represents the absolute value of the difference between that rater's proportion correct estimate for one of the 200 FPGEE standard-setting items, *i*, and the probability that a MQC would answer that item correctly. Consequently, instead of having a single summary fit index, each of the 24 raters included in the analysis had 200 fit indices (one for each item), which became the outcome variable for the two-level model I subsequently employed to analyze the data. Using this model, I was able to determine to what extent the differences in the rater fit indices (i.e., measures of the variance of the raters' proportion correct estimates) were related to item characteristic-related variables (i.e., each item's difficulty and

content domain classification) and rater background-related variables (i.e., each rater's content domain expertise and gender).

2.6.2 Item Characteristic-related Predictor Variables in the Level-1 Model

My purpose in specifying the level-1 model was to determine to what extent the content domain classifications of the items and the difficulty classifications of the items (i.e., two predictor variables) could account for the variance in the rater fit indices (i.e., the outcome variable).

For the purpose of the discussion below, let items be indexed by i = 1,…,I, and raters by r = 1,...,R. The level-1 predictor model has the form:

$$VARIANCE_{ir} = \beta_{0r} + \beta_{1r}CONTENT\_S_{ir} + \beta_{2r}CONTENT\_C_{ir} + \beta_{3r}DIFCAT\_ME_{ir} + \beta_{4r}DIFCAT\_MD_{ir} + \beta_{5r}DIFCAT\_D_{ir} + e_{ir}, \quad (7)$$

where the outcome variable $VARIANCE_{ir}$ represents the fit index associated with each item i $(i = 1, \dots, 200)$ and rater r $(r = 1, \dots, 24)$.

To ensure the full rank of the design matrix, I set the indicator variable for the BPS item content domain and the indicator variable for the *easy* item difficulty classification equal to 0. I chose the BPS item content domain classification as the reference category because there were more items (n = 93) classified in that content domain than in the other two content domains. I chose the *easy* item difficulty classification as the reference category for ease of interpretation as this classification was at the extreme end of the difficulty continuum (as was the *difficult* classification). It would have been more challenging to interpret results if I had chosen either the *moderately easy* or the *moderately difficult* classifications, since those two classifications flanked the mean *p*-value (+/- 1 SD).

Within this coding scheme, therefore, the intercept β_{0r} represents the mean fit index of the reference category (i.e., the mean fit index of rater *r* over all items classified as belonging to the BPS content domain classification and being of *easy* item difficulty). Furthermore, for each rater *r*, β_{1r} represents the mean difference between the fit index for items in the SAS content domain and those in the reference category; $CONTENT\_S_{ir}$ is the indicator variable associated with coefficient β_{1r}, with $CONTENT\_S_{ir} = 1$ if item *i* is a SAS content domain item and 0 otherwise, across all raters *r*; β_{2r} represents the mean difference between the fit index for items in the CS content domain and those in the reference category; $CONTENT\_C_{ir}$ is the indicator variable associated with coefficient β_{2r}, with $CONTENT\_C_{ir} = 1$ if item *i* is a CS content domain item and 0 otherwise, across all raters *r*; β_{3r} represents the mean difference between the fit index for *moderately easy* items and those in the reference category; $DIFCAT\_ME_{ir}$ is the indicator variable associated with coefficient β_{3r}, with $DIFCAT\_ME_{ir} = 1$ if item *i* is a *moderately easy* item and 0 otherwise, across all raters *r*; β_{4r} represents the mean difference between the fit index for *moderately difficult* items and those in the reference category; $DIFCAT\_MD_{ir}$ is the indicator variable associated with coefficient β_{4r}, with $DIFCAT\_MD_{ir} = 1$ if item *i* is a *moderately difficult* item and 0 otherwise, across all raters *r*; β_{5r} represents the mean difference between the fit index for *difficult* items and those in the reference category; and $DIFCAT\_D_{ir}$ is the indicator variable associated with coefficient β_{5r}, with $DIFCAT\_D_{ir} = 1$ if item *i* is a *difficult* item and 0 otherwise, across all raters *r*. The level-1 error term, $e_{ir} \sim \mathcal{N}(0, \sigma^2)$, assumes that errors are normally distributed with mean 0 and variance σ^2.

2.6.3 **Rater Background-related Predictor Variables in the Level-2 Model**

My purpose in specifying the level-2 model was to determine to what extent the raters' content domain expertise classifications and their gender classifications (i.e., predictor variables)

could explain the variance in their rater fit indices (i.e., outcome variable), after taking into account the variance in the rater fit indices that the item content domain classifications and the item difficulty classifications could explain (i.e., the level-1 predictor variables). The level-2 model includes covariates that are defined for each rater. Using this model allowed me to compare the contributions of these two rater background-related variables in explaining the variance in the rater fit indices.

I chose male as the reference category for gender since there were more male raters (*n* = 14) than female raters (*n* = 10) in my study sample. To maintain consistency in the interpretation of my results, I chose the BPS rater content domain expertise classification as the reference category since I had previously chosen BPS as the reference category for the item content domain classification.

The level-2 (rater-level) model has the following form:

$$\begin{aligned} \beta_{0r} &= \gamma_{00} + \gamma_{01} FEMALE_r + \gamma_{02} SAS_r + \gamma_{03} CS_r + u_{0r} \\ \beta_{1r} &= \gamma_{10} \\ \beta_{2r} &= \gamma_{20} \\ \beta_{3r} &= \gamma_{30} \\ \beta_{4r} &= \gamma_{40} \\ \beta_{5r} &= \gamma_{50} \end{aligned} \tag{8}$$

where β_{0r} is modeled as a random effect across raters and is dependent on the rater gender and rater content domain expertise classification; γ_{00} represents the average value of the fit index of the level-2 reference category (i.e., the fit index of a male rater of BPS content domain expertise); γ_{01} represents the mean difference between the fit index of a female rater and that of the level-2 reference category; $FEMALE_r$ is the indicator variable associated with coefficient

γ_{01}, with $FEMALE_r$= 1, if rater *r* is female, and 0 if male; γ_{02} represents the mean difference between the fit index of a rater with SAS content domain expertise and the level-2 reference category; SAS_r is the indicator variable associated with coefficient γ_{02}, with SAS_r= 1, if rater *r*'s content domain expertise is SAS, and 0 otherwise; γ_{03} represents the mean difference between the fit index of a rater with CS content domain expertise and the level-2 reference category; and CS_r is the indicator variable associated with coefficient γ_{03}, with CS_r= 1, if rater *r*'s content domain expertise is CS, and 0 otherwise. At level-2, coefficients β_{1r} through β_{5r} are modeled as fixed effects, γ_{10} through γ_{50}, respectively, indicating that each slope parameter described in the level-1 model remains constant within each rater *r*. The level-2 error term, $u_{0r} \sim \mathcal{N}(0, \tau_{00})$, is the random effect of slope β_{0r}, which represents the offset of each rater's fit index from the average γ_{00}.

I used HLM 7 Hierarchical Linear and Nonlinear Modeling (Scientific Software International, 2010) to conduct a two-level (items nested within raters design) analysis. The 24 raters provided proportion correct estimates for all 200 FPGEE items (i.e., there was no missing data in my design). I centered the reference categories at zero on the logit scale to set a baseline by which the relationships of the coefficients to the outcome variable were measured. The reference categories served as the baseline for comparing the relationships of the predictor variables in explaining the variance in the rater fit indices.

3. RESULTS

3.1 Summary of Results Addressing the Primary Research Question

In this study, I used hierarchical linear modeling to analyze a data structure in which items (level 1) were nested within raters (level 2). For the level-1 model, I included two item characteristic-related predictor variables: (a) item content domain classification, and (b) item difficulty classification. My goal was to determine to what extent the content domain classifications of the items and the difficulty classifications of the items could account for (i.e., help explain) the variance in the rater fit indices (i.e., the outcome variable). Overall, the two item characteristic-related variables accounted for 91% of the variance in the rater fit indices (i.e., variance that random effects did not account for).

See Table II (Appendix C) for the full results from my analysis. The regression coefficients relating the variance in the rater fit indices to the item content domain classifications of Social/Behavioral/Administrative Sciences ($\beta = 0.020$, $SE = 0.004$, $p < .001$) and Clinical Sciences ($\beta = 0.017$, $SE = 0.003$, $p < .001$) were both positive and significant. When judging the performance of MQCs on the 200 FPGEE items, the raters had more variation in their proportion correct estimates for items classified as Social/Behavioral/Administrative Sciences and Clinical Sciences items than for items classified as Basic Biomedical and Pharmaceutical Sciences items (i.e., the reference category).

The regression coefficients relating the variance in the rater fit indices to the item difficulty classifications of *moderately easy* ($\beta = 0.019$, $SE = 0.004$, $p < .001$), *moderately difficult* ($\beta = 0.024$, $SE = 0.004$, $p < .001$), and *difficult* ($\beta = 0.024$, $SE = 0.005$, $p < .001$) were all positive and statistically significant. When judging the performance of MQCs on the 200 FPGEE

items, the raters had more variation in their proportion correct estimates for items classified as *moderately easy*, *moderately difficult*, or *difficult* than for items classified as *easy* (i.e., the reference category), with the *moderately difficult* and *difficult* item difficulty classifications explaining somewhat more of the variance in the rater fit indices than the *moderately easy* item difficulty classification.

For the level-2 model, I added two rater background-related variables as predictor variables: (a) rater content domain expertise, and (b) rater gender. My goal was to determine how much of the variance in the rater fit indices was attributable to these two variables, after taking into account the variance in the rater fit indices that the item content domain classifications and the item difficulty classifications could explain (i.e., the level-1 predictor variables).

The regression coefficient relating the variance in the rater fit indices to rater gender was not statistically significant ($\beta = 0.000$, $SE = 0.014$, $p = .994$). Additionally, the regression coefficients relating the variance in the rater fit indices to rater content domain expertise in Clinical Sciences ($\beta = 0.007$, $SE = 0.017$, $p = .669$) and Social/Behavioral/Administrative Sciences ($\beta = -0.007$, $SE = 0.014$, $p = .624$) were not statistically significant. Apparently, these two rater background-related variables did not serve to moderate the association between the two item characteristic-related variables and the outcome variable. That is, the rater fit indices did not differ after grouping raters either by their gender classifications or by their content domain expertise classifications. After controlling for an item's content domain classification and difficulty classification, a rater's ability to provide an accurate estimate of the proportion of MQCs who would answer that item correctly was not related to the rater's gender or content domain expertise.

4. DISCUSSION AND CONCLUSIONS

The goal for this study was to determine to what extent four predictor variables (i.e., item content domain and difficulty classifications, and rater content domain expertise and gender classifications) could account for the variance in the rater fit indices. In this chapter, I summarize the outcomes of this study, identify potential applications and limitations of the study, address secondary research questions, and propose directions for future research.

4.1 Summary of Outcomes and Implications

Experts who advise organizations on the planning and conduct of standard settings recommend that the panelists (raters) should represent the population of practitioners (or experts) in the field. They advise organizations to design training programs to help raters understand the purpose of standard setting and their role in that process. Raters who are new to standard setting need opportunities to practice the tasks that they will perform, so experts recommend that the training programs should include practice rounds to help the raters become comfortable and proficient in performing the tasks. Skorupski and Hambleton (2005) reported that the accuracy of raters' judgments tended to improve after they participated in successive rounds in which they provided proportion correct estimates of candidate performance. During those practice rounds, the raters discussed candidate performance on individual items and got feedback on the accuracy of their initial estimates. However, even after participating in well-designed training programs that offer opportunities for practice, raters generally find it challenging to provide accurate proportion correct estimates of candidate performance on test items, particularly for MQCs (Brandon, 2004; Impara & Plake, 1998).

In this study, I proposed methods to determine the extent to which rater background-related variables and item characteristic-related variables could account for the variance in rater fit indices. I examined two rater background-related variables (i.e., gender and content domain expertise). These two variables explained little of the variance in the rater fit indices, after taking into account the variance in the rater fit indices that the content domain classifications and the item difficulty classifications explained. My results suggest that a rater's ability to provide an accurate estimate of the proportion of MQCs who would answer a FPGEE item correctly was not related to that rater's gender or content domain expertise. That is, the proportion correct estimates that female raters provided were no more (or less) accurate than the proportion correct estimates that male raters provided. Additionally, when I classified the raters according to their content domain expertise, the proportion correct estimates that the three groups of raters provided did not differ significantly in terms of their accuracy.

One might speculate that the rater selection process and the raters' participation in the intensive three-hour standard-setting training session may help explain these results. NABP considered the 24 practitioners whom they selected to participate in the FPGEE standard setting to be highly qualified to perform this task. They were all academicians from U.S. pharmacy programs with primary teaching responsibilities in one of the four core science domains taught in pharmacy schools. In the training session, raters engaged in discussions to help them reach a common understanding of what minimally qualified, yet sufficiently knowledgeable, U.S. pharmacy students who had completed the didactic portion of their education would know. They participated in several practice rounds, providing estimates of the proportion of minimally qualified U.S. pharmacy students who would answer each item correctly. In addition, the raters received feedback on the accuracy of their initial estimates and considered how their estimates

would impact the projected pass rate if NABP set the performance standard based solely on the outcomes of the standard setting. The discussions, participation in practice rounds, and feedback likely contributed to decreasing the variance in the raters' proportion correct estimates from one practice round to the next, even though the raters came to the standard-setting task with expertise in different content domains in pharmacy education.

I also examined two item characteristic-related variables (i.e., content domain classification and difficulty classification). Overall, 91% of the variance in the rater fit indices was attributable to the items' content domain classifications and difficulty classifications. When judging the performance of MQCs on the 200 FPGEE items, the raters exhibited more variance in proportion correct estimates for items classified as Social/Behavioral/Administrative Sciences and Clinical Sciences items than for items classified as Basic Biomedical and Pharmaceutical Sciences items (i.e., the reference category). Additionally, the raters had more variance in their proportion correct estimates for items classified as *moderately easy*, *moderately difficult*, or *difficult* than for items classified as *easy* (i.e., the reference category), with the *moderately difficult* and *difficult* items explaining somewhat more of the variance in the rater fit indices than the *moderately easy* items.

One must keep in mind that I calculated the rater fit indices used in this study from the absolute values of the differences in the raters' proportion correct estimates and empirical data. The values of those fit indices were restricted in range (i.e., 0.000 to 0.651), and the degrees of freedom for the item characteristic-related variables were large (i.e., 4,771), which had an impact on the study's statistical power. In other words, it is likely that even small differences between those rater fit indices that were associated with the various classifications of my item-level predictors would have been statistically significant. Therefore, while the results I reported for my

item characteristic-related predictor variables were statistically significant, it remains to be seen whether those results are practically significant.

Some researchers have reported that raters tend to overestimate the proportion of MQCs who would answer a difficult item correctly, while they underestimate the proportion of MQCs who would answer an easy item correctly (Clauser et al., 2009; Shepard, 1994; Taube, 1997). However, other researchers (Impara & Plake, 1998) have failed to replicate those findings, arguing that there is no significant relationship between raters' proportion correct estimates and the difficulties of items. My results provide insights into characteristics of FPGEE items that may make it more challenging for raters to estimate accurately the proportion of MQCs who would answer a given item correctly. However, the application of these insights in practice will need to be considered carefully, since the practical significance of my study's findings has yet to be determined.

When considering this study's results, one must keep in mind the constraints that NABP test developers face as they create a test form. The FPGEE test blueprint states that a test form must represent all four content domains. The blueprint also indicates how many items in each of the content domains a test form must include. Therefore, it would not be possible to exclude certain content domains from the FPGEE or to change the number of items included from each domain without first changing the test blueprint.

NABP should continue to include items that vary in difficulty in order to mirror the assembly of operational forms for candidates and also plan to build in sufficient time for raters to practice providing proportion correct estimates and for raters to receive feedback on the accuracy of their initial estimates. The results of my study suggest that items classified as *moderately easy*, *moderately difficult*, and *difficult* introduced more variance in the raters' proportion correct

estimates than items classified as *easy*. It may be worthwhile to provide additional practice rounds for the raters that include items trending "difficult" for the MQC. When preparing for a standard-setting training session, NABP could create additional examples of these types of items to give raters more opportunities for practice.

4.2 Addressing the Secondary Research Questions

1. **Is there justification for adjusting a rater's proportion correct estimates (or eliminating a rater's estimates) if a rater's fit index calculated from those estimates exceeds a certain upper limit? If so, what should that upper limit be? How would adjustments be made?**

Testing organizations can evaluate rater "fit" to determine whether they should take into account the proportion correct estimates that all of the raters provide when determining a cut score for a test, or whether they should adjust (or, alternatively, exclude) the estimates of inaccurate, misfitting raters. Not all raters are able to provide accurate estimates; some raters provide more accurate estimates than other raters. Testing organizations should consider developing a policy that they could use to make clear under what circumstances they would eliminate raters' inaccurate proportion correct estimates when setting cut scores. The policy would need to establish defensible criteria that the testing organization would use to identify inaccurate raters. Those criteria might include calculating and evaluating a fit index for each rater (in addition to evaluating more traditional classical test theory-based statistics such as the distribution (spread) of each rater's proportion correct estimates, a comparison of the mean of a rater's proportion correct estimates to the mean of all the other raters' proportion correct

estimates, a comparison of the standard error of a rater's proportion correct estimates to the standard error of all the other raters' proportion correct estimates).

2. **In practice, how could knowledge of rater background-related variables and/or item characteristic-related variables that can explain the variance in rater fit indices inform decisions regarding the selection of experts to serve on standard-setting panels, training models used in standard settings, and standard-setting outcomes? In other words, if a licensing or certification organization could establish that experts having certain background characteristics had rater fit indices that exceeded some predefined limit, should the organization consider that information when deciding whom to invite to serve on a standard-setting panel? If a licensing and certification organization could establish that one or more item characteristic-related variables explained much of the variance in the raters' fit indices, how could the organization use that information to inform the design of rater training?**

The results of this small study of the NABP standard-setting process indicate that differences in rater gender and content domain expertise explained little of the variance in the raters' fit indices. The ability to provide accurate proportion correct estimates did not appear to be related to a rater's gender or content domain expertise, which ought to be reassuring to NABP. However, the raters did differ in terms of the accuracy of their estimates.

In the future, licensing and certification organizations might consider using rater fit indices to help them identify subject matter experts who are best qualified to participate in standard settings. For example, it may be valuable to engage prospective raters in a simulation exercise, asking them to provide estimates of the proportion of MQCs who are likely to answer items correctly, compute rater fit indices based on their estimates, use those rater fit indices to

identify the more accurate raters, and then invite those raters to take part in the organization's standard-setting process. Use of this pre-screening process might decrease the amount of time needed to train subject matter experts to perform the standard-setting task and might result in the setting of more appropriate standards for examinations.

This study's findings would seem to confirm the results from Skorupski and Hambleton's (2005) study: providing opportunities during rater training for interaction and discussion leads to less variability and more consensus in their proportion correct estimates. Perhaps if NABP had used a different rater training model, or if the raters had not participated in multiple practice rounds, the outcomes of this study may have been different. As van der Linden (1982) suggested, researchers who are interested in studying the effectiveness of various standard-setting training models might consider comparing models using the rater fit indices as measures of rater accuracy. For example, researchers could compare models that incorporate different sets of activities (e.g., providing feedback to raters between practice rounds about the accuracy of their proportion correct estimates, encouraging discussions among raters about items that are harder to judge accurately, providing raters with impact data so that they could see the impact of applying their performance standard on a cohort of candidates). Researchers would then be in a better position to identify those training activities that facilitate better rater fit.

3. **Could the gathering and reporting of rater fit-related evidence contribute to the development of a validity argument to support the interpretations of test scores for a specified use? If so, how?**

Test developers who are providing validity evidence to support the interpretations of test scores for a specified use must document the processes they used to establish a cut score, which

functions as the operationalization of a performance standard (Haertel & Lorie, 2000). As the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) explain,

> Defining the minimum level of knowledge and skill required for licensure or certification is one of the most important and difficult tasks facing those responsible for credentialing. The validity of the interpretation of the test scores depends on whether the standard for passing makes an appropriate distinction between adequate and inadequate performance…. Verifying the appropriateness of the cut score or scores on a test used for licensure or certification is a critical element of the validation process. (p. 176)

Therefore, an important part of building a scientifically sound and convincing validity argument to support the use of test scores involves clearly describing the rationale and procedures that a testing organization used to set cut scores (Standard 5.21, p. 107). The *Standards* suggest that as part of the documentation of these procedures, test developers should report statistics that indicate the degree of variability in the judgments of the standard-setting participants (p. 108). The reporting of rater fit indices could provide that needed type of validity evidence.

Studies such as the one that I conducted provide validity evidence for test developers to consider as they review the processes that they are using for selecting raters and rater training models. The results of this study could be included in NABP's documentation of the FPGEE standard setting and in proposals for improving future standard-setting processes.

4.3 **Limitations**

This study of selected rater background-related variables and item characteristics-related variables has produced useful, practical information about standard-setting processes. However, the study had several limitations.

First, I used as data the proportion correct estimates that subject matter experts in pharmacy education assigned. Their backgrounds (i.e., their areas of content domain expertise) mirrored the over-arching content domains represented on the test. Other licensing and certification organizations may select subject matter experts to serve on standard-setting panels who are "generalists" in their fields, thus making it challenging to duplicate the conditions of this study.

When I conducted my research, I decided not to look at interactions of level-1 and level-2 variables, a limitation of this study. In future research, it would be interesting to investigate potential interactions of the item content domain classifications and the rater content domain expertise classifications. If researchers were to use the same model that I employed to study these interactions, they would need to add the rater content domain expertise classifications as the level-2 covariates for the β_1 and β_2 slope parameters.

Additional limitations of this study are the rater sample size and the number of items employed. I studied the outcomes of a single standard-setting exercise involving 24 raters and their proportion correct estimates for 200 items. It would be prudent to attempt to replicate this study under the same conditions (i.e., rater selection, training, etc.) for the same examination (FPGEE) program to see whether my results are generalizable beyond this one standard setting. Additionally, because the standard-setting process I studied involved a single organization's model for one examination, the results may not be generalizable to other licensing or certification programs using similar models.

The study limitations also include my treatment of item difficulty as a categorical variable. In future investigations of the relationship between item difficulty and the rater fit index, researchers might consider treating item difficulty as a continuous variable. Additionally,

I studied only two rater background-related variables. There are other rater background-related variables that may be of interest to researchers who are studying potential predictors of rater fit (e.g., rater race/ethnicity, years of teaching experience, and degree of prior involvement in the standard-setting process).

Finally, I investigated rater fit within the context of raters who participated in a modified Angoff standard-setting process. The results of this study may not generalize to the performance of raters who participate in other standard-setting processes. Researchers who are interested in this topic might consider studies addressing rater fit in the application of other standard-setting models.

4.4 **Conclusions**

In this study, I looked at several rater background-related and item characteristic-related variables in an attempt to determine to what extent those variables could explain the variance in raters' fit indices. While differences in rater gender and content domain expertise explained little of the variance, differences in the difficulties of the items and their content domain classifications held more explanatory power. The raters exhibited more variance in proportion correct estimates for items classified as Social/Behavioral/Administrative Sciences and Clinical Sciences items than for items classified as Basic Biomedical and Pharmaceutical Sciences items. Additionally, the raters had more variance in their proportion correct estimates for items classified as *moderately easy*, *moderately difficult*, or *difficult* than for items classified as *easy*, with the *moderately difficult* and *difficult* items explaining somewhat more of the variance in the rater fit indices than the *moderately easy* items.

The study's findings support the standard-setting experts' view that rater training that includes multiple practice rounds, discussions, interactions, and feedback can be influential in decreasing the variance in raters' estimates of the proportion of MQCs who would answer an item correctly. Additionally, the study's findings reinforce the importance of providing plenty of opportunities during training for raters to make judgments about candidate performance on items from different content domains, as well as items that differ in their levels of difficulty. Raters in this study had a more difficult time making accurate judgments about candidate performance on certain types of FPGEE items than on other types. NABP can use the results from the study to revise its rater training model to better prepare subject matter experts for their participation in FPGEE standard-setting processes.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 99-109). Washington, DC: American Council on Education.

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

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

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,* 1-21.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L Brennan (Ed.), *Educational measurement* (4th ed., pp. 701-731). Westport, CT: Praeger.

DeCoster, J., Gallucci, M., & Iselin, A.R. (2011). Best practices for using median splits, artificial categorization, and their continuous alternatives. *Journal of Experimental Psychopathology, 2,* 198-209.

Equal Employment Opportunity Commission. (1978). *Uniform Guidelines on Employee Selection Procedures,* 43 FR 38295.

Ferdous, A. A., & Plake, B. S. (2008). Item response theory-based approaches for computing minimum passing scores from an Angoff-based standard-setting study. *Educational and Psychological Measurement, 68*, 778-795.

Geisinger, K. F., & McCormick, C. M. (2010) Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29*(1), 38-44.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education, 12*, 13-28.

Haertel, E. H, & Lorie, W. A. (2000). *Validating standards-based test score interpretations*. Stanford, CA: Stanford University. Retrieved from http://statweb.stanford.edu/~rag/ed351/Std-Setting.pdf

Haladyna, T. M. (1994). A research agenda for licensing and certification testing validation studies. *Evaluation & the Health Professions, 17*(2), 242-256.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Linn (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Academic.

Hierarchical Linear and Nonlinear Modeling (HLM) (Version 7) [Computer software]. Skokie, IL: Scientific Software International.

.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63,* 594-601.

Hurtz, G. M., & Jones, J. P. (2009). Innovations in measuring rater accuracy in standard setting: Assessing "fit" to item characteristic curves. *Applied Measurement in Education, 22,* 120-143.

Hurtz, G. M., Jones, J. P., & Jones, C. N. (2008). Conversion of proportion-correct standard-setting judgments to cutoff scores on the IRT θ scale. *Applied Psychological Measurement, 32,* 385-406.

Impara, J. C., & Plake, B. S. (1996, April). *Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353-366.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69-80.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). Englewood Cliffs, NJ: Prentice-Hall.

Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice, 10*(2), 3-14.

Jaeger, R. M. (1995). On the cognitive construction of standard-setting judgment: The case of configural scoring. In *Proceedings of the Joint Conference on Standard Setting for*

Large-Scale Assessments (Vol. 22, pp. 57-73). Washington, DC: U.S. Government Printing Office.

Kane, M. T. (1987). On the use of IRT models with judgmental standard-setting procedures. *Journal of Educational Measurement, 24,* 333-345.

Kane, M. T. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, *64*, 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Meara, K. P., Hambleton, R. K., & Sireci, S. G., (2001). Setting and validating standards on professional licensure and certification exams: A survey of current practices. *CLEAR Exam Review, 12*(2), 17-23.

National Association of Boards of Pharmacy. (n.d.). *FPGEE Competency Statements*. Retrieved from http://www.nabp.net/programs/examination/fpgee/fpgee-blueprint.

Norcini, J. J. (1994). Research on standards for professional license and certification examinations. *Evaluation & Health Professions, 17,* 160-177.

Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement, 25*(1), 57-65.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education, 1,* 65-80.

Plake, B. S., Impara, J. C., & Potenza, M. T. (1994). Content specificity of expert judgments in a standard-setting study. *Journal of Educational Measurement, 31*(4), 339-347.

Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement, 28,* 249-256.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago, IL: University of Chicago Press.

Raymond, M. R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 181-223). Mahwah, NJ: Lawrence Erlbaum Associates.

Reckase, M. D. (2001). Innovative methods for helping standard setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. Cizek (Ed.), *Standard setting: Concepts, methods and perspectives* (pp. 159-173). Mahwah, NJ: Erlbaum.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10*(2), 11-14.

Shepard, L. A. (1994). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*. Washington, DC: National Assessment Governing Board and National Center for Educational Statistics.

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, innovations* (pp. 135-148). New York, NY: Routledge.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 19*(3), 233-256.

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

Taub, K. T. (1997). The incorporation of empirical item difficulty data into the Angoff standard-setting procedure. *Evaluation and the Health Professions, 20*(4), 479-498.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press

APPENDICES

APPENDIX A[1]

FPGEE Competency Statements

Area 1 - Basic Biomedical Sciences

1A Physiology
- 1A01 structure and function of major body systems; as it applies to integumentary, muscular skeletal, cardiovascular, lymphatic, respiratory, digestive, nervous, endocrine, urinary, reproductive, and body fluids and electrolytes, cells in tissue

1B Biochemistry
- 1B01 chemistry of biomacromolecules (proteins, lipids, carbohydrates, and DNA)
- 1B02 nucleic acid biosynthesis and metabolism
- 1B03 enzymology and coenzymes and kinetics
- 1B04 metabolic pathways to energy utilization

1C Microbiology
- 1C01 general principles of microbial concepts
- 1C02 principles of infectious diseases
- 1C03 host-parasite relationships
- 1C04 pathogenic microorganisms of man
- 1C05 inflammatory responses to infectious agents

1D Molecular Cell Biology/Genetics
- 1D01 gene expression
- 1D02 carrier proteins/membrane transport
- 1D03 mechanics of cell division
- 1D04 ion channels and receptor physiology
- 1D05 chromosomes and DNA
- 1D06 gene transcription and translation processes
- 1D07 recombinant DNA technology

1E Immunology
- 1E01 human immunity and immune responses
- 1E02 principles of antigen-antibody relationships
- 1E03 antibody synthesis, development, function and immunopathology

Area 2 - Pharmaceutical Sciences

2A Medicinal Chemistry
- 2A01 physiochemical properties of drugs in relation to drug absorption, distribution, metabolism, and excretion (ADME)
- 2A02 chemical basis for drug action
- 2A03 fundamental pharmacophores for drugs used to treat diseases
- 2A04 structure activity relationships in relation to drug-target interactions
- 2A05 chemical pathways of drug metabolism
- 2A06 applicability to making drug therapy decisions

2B Pharmacology and Toxicology
- 2B01 mechanisms of action of drugs of various categories
- 2B02 pharmacodynamics of drug action and absorption, distribution, metabolism, and elimination
- 2B03 adverse effects and side-effects of drugs
- 2B04 drug-target interactions
- 2B05 drug discovery and development
- 2B06 mechanism of toxicity and toxicokinetics
- 2B07 acute and chronic toxic effect of xenobiotics, including drug and chemical overdose and toxic signs of drugs of abuse
- 2B08 interpretation of drug screens
- 2B09 principles of antidotes and alternative approaches to toxic exposures
- 2B10 functions of poison control centers
- 2B11 bioterrorism and disaster preparedness and management

2C Pharmacognosy and Alternative and Complementary Treatments
- 2C01 concepts of crude drugs, semi-purified, and purified natural products
- 2C02 evaluation of alternative and complementary medicine purity, bioavailability, safety, and efficacy
- 2C03 classes of pharmacologically active natural products
- 2C04 Science of dietary supplements (vitamins, minerals, and herbals)
- 2C05 Dietary Health Supplement and Education Act and Impact on regulation of dietary supplements and herbal products

2D Pharmaceutics
- 2D01 physiochemical principles of dosage forms
- 2D02 principles of drug delivery via dosage forms (eg, liquid, solid, semi-solid, controlled release, patches, and implants)
- 2D03 principles of dosage form stability and drug degradation in dosage forms
- 2D04 materials and methods used in preparation and use of drug forms

2E Biopharmaceutics/Pharmacokinetics
- 2E01 biological principles of dosage forms
- 2E02 basic principles of in vivo drug kinetics (linear and nonlinear)
- 2E03 principles of bioavailability/bioequivalence
- 2E04 physiologic determinates of drug onset and duration
- 2E05 drug, disease, and dietary influences on absorption, distribution, metabolism, and excretion
- 2E06 the pharmacokinetic-pharmacodynamic interface

2F Pharmacogenomics
- 2F01 genetic basis for disease and drug action
- 2F02 genetic basis for alteration and drug metabolism
- 2F03 genome and proteomic principles in relation to disease and drug development
- 2F04 genetic basis for individualizing drug doses

2G Extemporaneous Compounding/Parenteral/Enteral
- 2G01 United States Pharmacopeia guidance on compounding and FDA Compliance Policy Guidelines
- 2G02 techniques and principles used to prepare and dispense individual extemporaneous prescriptions including dating of compounded dosage forms
- 2G03 extemporaneous liquid (parenteral, enteral), solid, semi-solid, and topical preparations
- 2G04 dosage form preparation calculations
- 2G05 sterile admixture techniques
  - a United States Pharmacopeia (USP) Chapter <797>
  - b stability and sterility testing and dating
  - c clean room requirements
  - d infusion devices and catheters

Area 3 - Social/Behavioral/Administrative Sciences

3A Health Care and Public Health Delivery Systems
- 3A01 introduction to United States, state, and local health care delivery systems and their interfaces and how they compare to those in other industrialized countries
- 3A02 social, political, and economic factors influencing the delivery of health care (including financing and reimbursement mechanisms, health disparities, reform, etc)
- 3A03 pharmacy and health care organizations (private and public insurers of third party administration, pharmaceutical industry, managed care organizations, PBMs, etc)
- 3A04 health policy development and evaluation
- 3A05 importance of involvement in pharmacy organizational, regulatory, state, and federal issues
- 3A06 conflict between medical care and public health
- 3A07 contributions of public health efforts to health status improvements (infectious disease control, chronic disease preventions, demographics, and social and physical environmental factors, etc)

3B Economics/Pharmacoeconomics
- 3B01 use of pharmacoeconomic analyses (ie, cost-benefit analysis, cost-effectiveness analysis, cost-minimization analysis, cost-utility analysis)
- 3B02 applications of economic, clinical, and humanistic outcomes to improve allocation of limited health care resources

3B03 general macro and micro economic principles

3C Pharmacy Management

- 3C01 management principles (planning, organizing, directing, and controlling pharmacy resources) applied to various pharmacy practice setting and patient outcomes
- 3C02 personnel management - including leadership
- 3C03 managing goods and services (marketing, purchasing/inventory management, and merchandising)
- 3C04 financial accounting
- 3C05 risk management in pharmacy practice

3D Pharmacoepidemiology

- 3D01 application of epidemiological study designs to study drug use and outcomes in large populations
- 3D02 data sources and analytic tools that provide an estimate of the probability of beneficial or adverse effects of medication use in large populations
- 3D03 methods for continually monitoring unwanted effects and other safety-related aspects of medication use in large populations

3E Pharmacy Law and Regulatory Affairs

- 3E01 administrative, civil, and criminal liability
- 3E02 a pharmacist's responsibilities and limits under the law
- 3E03 the authority, responsibilities, and operation of agencies and entities that administer laws and regulations related to prescription, and over-the-counter medications

3F Biostatistics and Research Design

- 3F01 commonly used experimental and observational study designs
- 3F02 commonly used statistical tests and their appropriate application
- 3F03 evaluation of statistical results including an understanding of statistical versus clinical significance

3G Ethics

- 3G01 principles of biomedical ethics
- 3G02 ethical dilemmas in the delivery of patient-centered care, including
  - a conflicts of interest
  - b end-of-life decision making
  - c development, promotion, sales, prescription, and use of drugs
  - d working in groups
- 3G03 research ethics
- 3G04 professional behavior (ie, professionalism, code of ethics, oath of the pharmacist)

3H Core Communication Concepts and Skills

- 3H01 patient counseling skills including active listening and empathy
- 3H02 assertiveness and problem-solving techniques, handling difficult situations - patients and other core providers
- 3H03 interviewing techniques
- 3H04 health literacy
- 3H05 cultural competency

3I Social and Behavioral Aspects encountered in Practice
- 3I01 health, illness, and sick role behaviors
- 3I02 principles of behavior modification
- 3I03 patient adherence
- 3I04 caregiving throughout the life cycle
- 3I05 death and dying
- 3I06 patients' and other health care providers' perceptions of pharmacists' capabilities

3J Medication Dispensing and Distribution Systems
- 3J01 safe and effective preparation and dispensing of medications in all types of practice settings
- 3J02 development and maintenance of patient medication profiles
- 3J03 role of automation and technology
- 3J04 continuous quality improvement programs or protocols in the medication-use process, including identification and prevention of medication errors and establishment of error reduction programs, technology of drug information retrieval for quality assurance

Area 4 - Clinical Sciences

4A Literature Evaluation - Practice Guidelines and Clinical Trials
- 4A01 principles of clinical practice guidelines for various disease states and their interpretation in the clinical setting
- 4A02 integration of core scientific and systems-based knowledge in patient care decisions
- 4A03 reinforcement of basic science principles relative to drug treatment protocols and clinical practice guidelines
- 4A04 evaluation of clinical trials that validate treatment usefulness

4B Drug Information
- 4B01 fundamentals of the practice of drug information
- 4B02 application of drug information skills for delivery of medication therapy management
- 4B03 the ability to judge the reliability of various sources of information

4C Clinical Pathophysiology
- 4C01 pathophysiology of disease states amenable to pharmacist intervention

4D Clinical Pharmacokinetics/Pharmacogenomics
- 4D01 clinical pharmacokinetics/pharmacogenomics of commonly used and low-therapeutic-index drugs
- 4D02 clinical basis for individualizing drug therapy

4E Clinical Prevention and Population Health
- 4E01 promotion of wellness and nonpharmacologic therapies
- 4E02 disease prevention and monitoring

4F Medication Therapy Management - Patient Assessment, Clinical Pharmacology, and Therapeutics
- 4F01 concepts of pharmacist-provided patient care and medication therapy management services
- 4F02 importance of and techniques for obtaining a comprehensive patient history
- 4F03 patient assessment (e.g., inspection, palpation, percussion, auscultation), terminology, and the modifications caused by common disease states and drug therapy

4F04	common clinical laboratory values and diagnostic tests and their clinical role
4F05	OTC point-of-care testing devices (e.g., glucometers, pregnancy tests, home testing for HbA1c, drug screening).
4F06	false positive and false negative results
4F07	therapeutic drug concentrations and their interpretation
4F08	problem identification (e.g., duplication dosage, drug interactions, dietary interactions, adverse drug reactions and interactions, frequency dosage form, indication mismatches) and resolution planning
4F09	triage and referral skills
4F10	designing of patient-centered, culturally relevant treatment plans
4F11	application of evidence-based decision making to patient care
4F12	nonprescription and dietary supplements
4F13	drug monitoring for positive and negative outcomes (including drug induced disease)
4F14	clinical management of drug toxicity and overdo

[1] The *FPGEE Competency Statements* may be accessed at the National Association of Boards of Pharmacy website: http://www.nabp.net/programs/examination/fpgee/fpgee-blueprint

APPENDIX B

Table I

Rater Background-Related Characteristics and Errors of Specification (E_{Ir})

Rater	Content Domain Expertise	Gender	E_{Ir} (mean variance over all 200 items)
1	SAS	F	0.170
2	SAS	M	0.279
3	BPS	M	0.211
4	CS	F	0.142
5	CS	M	0.139
6	CS	M	0.144
7	SAS	M	0.141
8	SAS	F	0.141
9	CS	F	0.180
10	CS	F	0.148
11	BPS	M	0.129
12	SAS	M	0.113
13	BPS	M	0.134
14	CS	F	0.140
15	CS	M	0.136
16	CS	F	0.161
17	BPS	F	0.135
18	CS	M	0.135
19	SAS	M	0.154
20	CS	M	0.150
21	BPS	M	0.126
22	CS	F	0.189
23	CS	F	0.128
24	BPS	M	0.139

APPENDIX C

Table II

Final Estimation of Fixed Effects (with Robust Standard Errors)

Fixed Effect	*β*	*SE*	*t*	*df*	*p*
β_{0r}	0.121	0.014	8.43	20	< 0.001
FEMALE$_r$, γ_{01}	0.000	0.014	0.01	20	0.994
SAS$_r$, γ_{02}	-0.007	0.014	-0.50	20	0.624
CS$_r$, γ_{03}	0.007	0.017	0.39	20	0.699
CONTENTS$_i$, β_{1r}	0.020	0.004	5.59	4771	< 0.001
CONTENTC$_i$,β_{2r}	0.017	0.003	6.01	4771	< 0.001
DIFCATME$_i$,β_{3r}	0.019	0.004	4.36	4771	< 0.001
DIFCATMD$_i$,β_{4r}	0.024	0.004	6.17	4771	< 0.001
DIFCATD$_i$, β_{5r}	0.024	0.005	4.77	4771	< 0.001

VITA

NAME:	Maria Incrocci
EDUCATION:	B.S., Pharmacy, University of Illinois, College of Pharmacy, Chicago, Illinois, 1984
	M.S., Biological Sciences, Chicago State University, Chicago, Illinois, 2005
	Ph.D., Educational Psychology, University of Illinois, Chicago, Illinois, 2015
EXPERIENCE:	Competency Assessment Senior Manager, National Association of Boards of Pharmacy, Mount Prospect, Illinois, 2007-present
	Competency Assessment Program Analyst, National Association of Boards of Pharmacy, Mount Prospect, Illinois, 2006-2007
	Pharmacy Manager, Pharmacist, Osco Drug, Franklin Park, Illinois, 1984-2008
TEACHING:	Departments of Health Sciences and Physical Education, Moraine Valley Community College, Palos Hills, Illinois, 1999-2007
PROFESSIONAL AFFILIATIONS:	American Pharmacists Association American Society of Health Systems Pharmacists National Council on Measurement in Education
LICENSES AND CERTIFICATIONS:	APhA Pharmacy–Based Immunization Delivery Certification Advanced Cardiac Life Support Certification Illinois Licensed Pharmacist
PUBLICATIONS:	Boyle, M., & Myford, C. (2013). Pharmacists' expectations for entry-level practitioner competency. *American Journal of Pharmaceutical Education, 77*(1), Article 5.
	Newton, D. W., Boyle, M., & Catizone, C.A. (2008). The NAPLEX: Evolution, purpose, scope and educational implications. *American Journal of Pharmaceutical Education, 72*(2): 33.

PAPER PRESENTATIONS: Newton, D. W., Boyle, M., & Catizone, C. A., (2008, July). *The NAPLEX: Evolution, purpose, scope and educational implications.* Paper presented at the annual meeting of the American Association of Colleges of Pharmacy, Chicago, IL.