University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Open Access Dissertations

9-2013

Examination of the Application of Item Response Theory to the Angoff Standard Setting Procedure

Jerome Cody Clauser
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/open_access_dissertations

Part of the Education Commons

Recommended Citation

Clauser, Jerome Cody, "Examination of the Application of Item Response Theory to the Angoff Standard Setting Procedure" (2013). *Open Access Dissertations*. 783.
https://doi.org/10.7275/432y-0b36 https://scholarworks.umass.edu/open_access_dissertations/783

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

EXAMINATION OF THE APPLICATION OF ITEM RESPONSE THEORY TO THE ANGOFF STANDARD SETTING PROCEDURE

A Dissertation Presented

By

JEROME CODY CLAUSER

Submitted to the Graduate School of the

University of Massachusetts Amherst in partial fulfillment

Of the requirements for the degree of

DOCTOR OF EDUCATION

September 2013

Education

© Copyright Jerome Cody Clauser 2013

All Rights Reserved

EXAMINATION OF THE APPLICATION OF ITEM RESPONSE THEORY TO THE ANGOFF STANDARD SETTING PROCEDURE

A Dissertation Presented

By

JEROME CODY CLAUSER

Approved as to style and content by:

Ronald K. Hambleton, Chairperson

Lisa A. Keller, Member

Penelope S. Pekow, Member

Christine B. McCormick, Dean

School of Education

ACKNOWLEDGMENTS

I am profoundly grateful and humbled by the encouragement and guidance I have received throughout the writing of this dissertation. First, I appreciate the support of my committee members, each of whom has made invaluable contributions throughout this process.

Penelope Pekow brought the perfect balance of expertise and perspective to the committee. Your careful reading, thoughtful comments, and excellent suggestions have dramatically improved this thesis.

Lisa Keller has been a friend and mentor for over four years. Your endless ability to both challenge and support has made me a better researcher, a better thinker, and a better person.

Finally, Ron Hambleton has been the greatest, advisor, teacher, and mentor that I ever could have imagined. Your wisdom and insight have been unparalleled as, together, we reexamined what we knew about Standard Setting. I am honored to be able to call you my teacher, and equally honored to call you my friend.

In addition to my committee members, I am tremendously thankful for the excellent community of researchers and scholars in the Psychometric Methods program. I feel particularly fortunate that my time at UMass has been shared with a cohort of amazingly gifted and thoughtful students. When I joined the program I did not get to select my peers, but if I had, I would have chosen each of you.

I would like to thank my partner, Amanda, who has supported me both emotionally and intellectually throughout this process. I am so thankful to have been able to share this journey with you. I could not have asked for a better partner in psychometrics nor in life.

Finally I would like to acknowledge the tremendous support I have received from my parents over the last four years. We have shared in this process in a way that few families could, and I feel truly blessed that you are now able to share in my success. Your love and support are truly appreciated.

ABSTRACT

EXAMINATION OF THE APPLICATION OF ITEM RESPONSE THEORY TO THE ANGOFF STANDARD SETTING PROCEDURE

SEPTEMBER 2013

JEROME CODY CLAUSER, B.S. WEST CHESTER UNIVERSITY OF PENNSYLVANIA

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Ronald K. Hambleton

Establishing valid and reliable passing scores is a vital activity for any examination used to make classification decisions. Although there are many different approaches to setting passing scores, this thesis is focused specifically on the Angoff standard setting method. The Angoff method is a test-centric classical test theory based approach to estimating performance standards. In the Angoff method each judge estimates the proportion of minimally competent examinees who will answer each item correctly. These values are summed across items and averages across judges to arrive at a recommended passing score. Unfortunately, research has shown that the Angoff method has a number of limitations which have the potential to undermine both the validity and reliability of the resulting standard.

Many of the limitations of the Angoff method can be linked to its grounding in classical test theory. The purpose of this study is to determine if the limitations of the Angoff could be mitigated by a transition to an item response theory (IRT) framework. Item response theory is a modern measurement model for relating examinees' latent ability to their observed test performance. Theoretically the transition to an IRT-based Angoff method could result in more accurate, stable, and efficient passing scores.

The methodology for the study was divided into three studies designed to assess the potential advantages of using an IRT-based Angoff method. Study one examined the effect of allowing judges to skip unfamiliar items during the ratings process. The goal of this study was to detect if passing scores are artificially biased due to deficits in the content experts' specific item level content knowledge. Study two explored the potential benefit of setting passing scores on an adaptively selected subset of test items. This study attempted to leverage IRT's score invariance property to more efficiently estimate passing scores. Finally study three compared IRT-based standards to traditional Angoff standards using a simulation study. The goal of this study was to determine if passing scores set using the IRT Angoff method had greater stability and accuracy than those set using the common True Score Angoff method. Together these three studies examined the potential advantages of an IRT-based approach to setting passing scores.

The results indicate that the IRT Angoff method does not produce more reliable passing score than the common Angoff method. The transition to the IRT-based approach, however, does effectively ameliorate two sources of systematic error in the common Angoff method. The first source of error is brought on by requiring that all judges rate all items and the second source is introduced during the transition from test to scaled score passing scores. By eliminating these sources of error the IRT-based method allows for accurate and unbiased estimation of the judges' true opinion of the ability of the minimally capable examinee.

Although all of the theoretical benefits of the IRT Angoff method could not be demonstrated empirically, the results of this thesis are extremely encouraging. The IRT Angoff method was shown to eliminate two sources of systematic error resulting in more accurate passing scores. In addition this thesis provides a strong foundation for a variety of studies with the potential to aid in the selection, training, and evaluation of content experts. Overall findings from this thesis suggest that the application of IRT to the Angoff standard setting method has the potential to offer significantly more valid passing scores.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS....iv

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER

INTRODUCTION 1

1.1 The Basics of Standard Setting.... 1
1.2 The Angoff Method.... 2
1.3 The Angoff Method on the IRT Scale 5
1.4 Statement of Problem 13
1.5 Statement of the Purpose 15
1.6 Outline of the Study 15

LITERATURE REVIEW 17

2.1 Overview of Literature Review 17
2.2 Conceptualizing the Minimally Capable Examinee.... 18
2.3 Internal Consistency of Judges' Ratings 21
2.4. Feedback Between Rounds 25
2.5. Placing Angoff Ratings onto the Score Scale.... 30
2.6 Setting Angoff Standards Using IRT.... 33
2.7. Selective and Adaptive Rating of Test Items 40
2.8 Summary of the Literature Review 44

SELECTIVE STANDARD SETTING.... 47

3.1 Background and Purpose 47
3.2 Research Question.... 48
3.3 Data 48
3.4 Methodology.... 50
3.5 Evaluating Results.... 51
3.6 Results.... 51

ADAPTIVE STANDARD SETTING 56

4.1 Background and Purpose ... 56
4.2 Research Questions ... 56
4.3 Data ... 57
4.4 Methodology ... 61
4.5 Evaluating Results ... 63
4.6 Results ... 64

STABILITY OF PERFORMANCE STANDARDS ... 68

5.1 Background and Purpose ... 68
5.2 Research Questions ... 69
5.3 Data ... 69
5.4 Methodology ... 70
5.5 Evaluating Results ... 75
5.6 Results ... 77

DISCUSSION ... 82

6.1 Introduction ... 82
6.2 Selective Standard Setting ... 84
6.3 Adaptive Standard Setting ... 90
6.4 Stability of Performance Standards ... 95
6.5 Overall Discussion ... 99
6.6 Limitations ... 103
6.7 Future Research ... 104

REFERENCES ... 107

LIST OF TABLES

Table Page

1 Comparison of Mean and Median Passing Score....................10

2 Position of Marked Ratings Step 1....................52

3 Position of Marked Ratings Step 2....................52

4 Step 1 Change in Recommended Passing Score....................54

5 Step 2 Change in Recommended Passing Score....................54

6 Range of Acceptable Passing Scores Step 1....................64

7 Range of Acceptable Passing Scores Step 2....................65

8 Adaptive Passing Scores Step 1....................65

9 Adaptive Passing Scores Step 2....................65

10 Probability of Observing More Accurate Passing Score Step 1....................66

11 Probability of Observing More Accurate Passing Score Step 2....................66

12 Standard Deviation for Passing Scores at Each Difficulty Condition....................77

13 Mean Passing Scores at Different Difficulties Conditions....................79

14 Mean Absolute Difference between True and Observed Passing Scores....................80

LIST OF FIGURES

Figure Page

1 Expected Angoff Ratings for the Internally Consistent Judge 6

2 Item Level Theta Estimates Bases on Angoff Ratings 8

3 Add Inconsistent Judge 10

4 Add Consistent Judge 10

5 Test Characteristic Curve 1 31

6 Test Characteristic Curve 2 31

7 TCC Comparison 32

8 Step 1 Panel 1: Test Information and Ratings Density 58

9 Step 1 Panel 2: Test Information and Ratings Density 58

10 Step 1 Panel 3: Test Information and Ratings Density 59

11 Step 2 Panel 1: Test Information and Ratings Density 59

12 Step 2 Panel 2: Test Information and Ratings Density 60

13 Step 2 Panel 3: Test Information and Ratings Density 60

14 Distribution of True Score Angoff Passing Scores 81

CHAPTER 1

INTRODUCTION

1.1 The Basics of Standard Setting

In criterion-referenced testing, examinees' performance is assessed in relation to a domain of content. When an examinee's test score results in a categorical decision, such as pass/fail or basic/proficient/advanced, expert judgment is required to determine what level of domain mastery is necessary for examinees to attain each performance level. The points on the score scale which separate these performance categories, known as performance standards, cut scores, or passing scores are not typically arrived at strictly through empirical analysis. Instead, experts familiar with both the examinee population for the test and the content domain provide judgments as to what level of content mastery is "minimally acceptable" or "just enough" to be placed in each performance category . The process of establishing cut scores, known as standard setting, is a systematic and typically iterative procedure for the placement of expert opinions on the score scale. Because passing scores are the product of expert judgment, there is no one true passing score to be discovered. Instead, standard setting procedures provide a systematic method for inferring passing scores from a diverse panel of content experts often influenced by empirical evidence (Reckase, 2000). These individual judgments are then combined through a variety of methods to arrive at a single recommended passing score (e.g., Cizek, 2001), or additional passing scores too, if that is the intent of the process.

Establishing performance standards is a fundamental part of the test development process for any examination used for the classification of individuals. Inappropriate passing scores can have far reaching negative consequences for both individuals and society at large. These consequences include everything from depriving a qualified student of a high school diploma to licensing a dangerously under-qualified physician. The validity of these passing scores is therefore fundamental to the integrity of any assessment used to make classification decisions.

Although standard setting has important implications for virtually all areas of testing, this study will focus specifically on standard setting on certification and licensure examinations. Unlike educational achievements tests where typically multiple performance standards are set, credentialing exams set a single passing score. Furthermore, the technical nature of content covered on credentialing exams may have unique implications for the standard setting process. Although many of the issues discussed below will be applicable to both credentialing and educational achievement tests, at times the results may not be generalizable.

1.2 The Angoff Method

The most popular and well researched standard setting procedure in use on credentialing exams today is the method originally described by Angoff in 1971 (Angoff, 1971, Meara, Hambleton, & Sireci, 2001).The Angoff method is a test-centered standard setting procedure, meaning that judgments are rendered about the test items rather than the individual examinees. Like all test centered methods, content experts participating begin by considering what it means for an examinee to be minimally acceptable. This theoretical minimally competent examinee (MCE) is one whose knowledge, skills, and abilities are just barely sufficient to qualify for a given performance category. Each judge relies on his or her experience with the content domain to conceptualize an examinee whose mastery is considered just barely acceptable (Livingston & Zieky, 1982). This notion is fundamental to the Angoff method, since the test score of the minimally acceptable examinee will be the judge’s recommended passing score.

After each content expert has developed his or her opinion of the capability of the minimally acceptable candidate, this judgment must be placed onto the test score scale. In the Angoff method each judge is asked to provide an estimate of the probability that the minimally acceptable candidate would answer each dichotomously-scored item correctly. These probabilities are summed across items to arrive at a recommended cut score for each judge on the raw score scale. These judge level cut scores are then averaged to arrive at a recommended cut score for the panel of judges. Although a variety of Angoff modifications are often used in practice (Plake, Melican, & Mills, 1991), this straightforward procedure is always the basis for inferring the appropriate position of the performance standard on the test score scale.

1.2.2 Limitations of the Angoff Method

Although the logic undergirding the Angoff method is straightforward and appealing, actual implementation of the method has often proven to be complicated (Shepard, 1995; National Research Council, 1999). The primary obstacle for the method has been the inability of judges on occasion to perform the required task. Research on this topic has shown that although the judges may have the requisite skills to conceptualize the minimally competent examinee (Giraud, Impara, & Plake, 2005), they often struggle to provide reasonable estimates of that examinee's performance on particular test items (Busch & Jaeger, 1990; Clauser, Harik, et al., 2009; Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Clauser, Mee, & Margolis, 2011; Clauser, Swanson, & Harik, 2002; Clauser, Clauser, & Hambleton, 2012). Although there is no absolute criterion for the accuracy of expert judgments, the internal consistency of these ratings has been seen as an important framework for evaluation (Kane, 2001). Lack of internal consistency is typically illustrated through discordance between judges' probability estimates and item difficulty. For example a judge who does not produce internally consistent ratings may estimate that 70% of the minimally proficient examinees will answer a difficult item correctly but that only 30% will correctly answer an easier item. Insofar as this limitation results in a shift in the recommended cut score, this method fails to reflect the judges' view of the minimally acceptable examinee.

In addition to the practical limitations in judges' ability to perform the required task, serious theoretical concerns persist regarding the applicability of the Angoff method to modern testing applications. The Angoff standard setting method conceptualizes performance standards within a classical test theory framework and as a result produces performance standards on the true score metric. Within the true score framework an examinee's observed ability is dependent on the specific set of items included in the test. This item dependent view of examinee ability means that the theoretical performance of the minimally acceptable examinee, and therefore the cut score, will be item dependent. In practice, the influence of item selection is mitigated by translating the cut score on the test score scale onto the IRT proficiency scale through the test characteristic curve (TCC), but this translation does not ensure a consistent passing score regardless of item selection (Ferdous & Plake, 2008).

Finally, defining the performance standard on the test score scale requires that content experts provide probability estimates for all test items, regardless of their familiarity or comfort with the item content. Because classical test theory provides test, rather than item, level measurement, all test items must be rated for the accurate placement of a passing score. Not only is this requirement time consuming, but it forces judges to sometimes rate items outside their area of expertise. Although content experts are presumably familiar with the vast majority of the tested material, gaps in a judge's content knowledge are often unavoidable. The prevalence of these content deficits will vary across tests but have the potential to be particularly influential in the event that content experts are selected to represent non-expert constituencies or the test assesses highly technical material. Regardless of the reason, when judges are asked to provide ratings for items they themselves cannot answer it is difficult to argue that the judgments reflect the ability of minimally competent examinees. If errors in the rating of these items are random, the impact on the final passing score may be trivial. If, however, judges interpret items with which they struggle as systematically more difficult, the passing score may be artificially low. By defining performance standards on the test score scale, the Angoff method may at times compel judges to provide fallacious ratings.

1.3 The Angoff Method on the IRT Scale

Many of these limitations can theoretically be mitigated by conceptualizing the Angoff method within an item response theory framework. Item response theory (IRT) is a measurement model for relating an examinee's latent ability to their test performance (Hambleton & Swainathan, 1985; Hambleton, Swainathan, & Rodgers 1991; Nering & Ostini, 2010). Within IRT, examinee ability for a given content domain is represented as a point along a unidimensional proficiency continuum referred to as "theta." Although this latent ability is unobservable, examinee ability is estimated based on observed item responses. This relationship between an examinee's latent ability and his/her response on a given item can be described by a monotonically increasing function known as an item characteristic curve (ICC). Item characteristic curves are s-shaped functions bounded between zero and one which represent the probability of a correct response on the given item for an examinee at any point along the ability continuum. These functions allow for the probabilistic estimation of an examinee's ability based on a given response pattern. Conceptualizing the Angoff method within an IRT framework does not require any change in the judgment process. Instead, an IRT Angoff method simply applies IRT concepts to the interpretation of traditional Angoff ratings.

In the Angoff method the ability of the minimally acceptable examinee exists as a theoretical concept, wholly separate from the underlying measurement model. Although judges are expected to internalize a consistent view of examinee ability throughout the rating process, there is no attempt to place this underlying latent trait on an ability scale. In the Angoff method within an IRT framework, the ability of the minimally competent examinee is viewed as a point along the proficiency scale. This is not to suggest that judges are familiar with the mechanics of IRT or the particular features of the underlying score scale. Instead it simply requires that the ability of the minimally acceptable examinee can exist along the same scale as the ability of all other examinees.

The theta score for the minimally acceptable examinee can be unique for each judge but is expected to be consistent for a single content expert across a round of ratings.

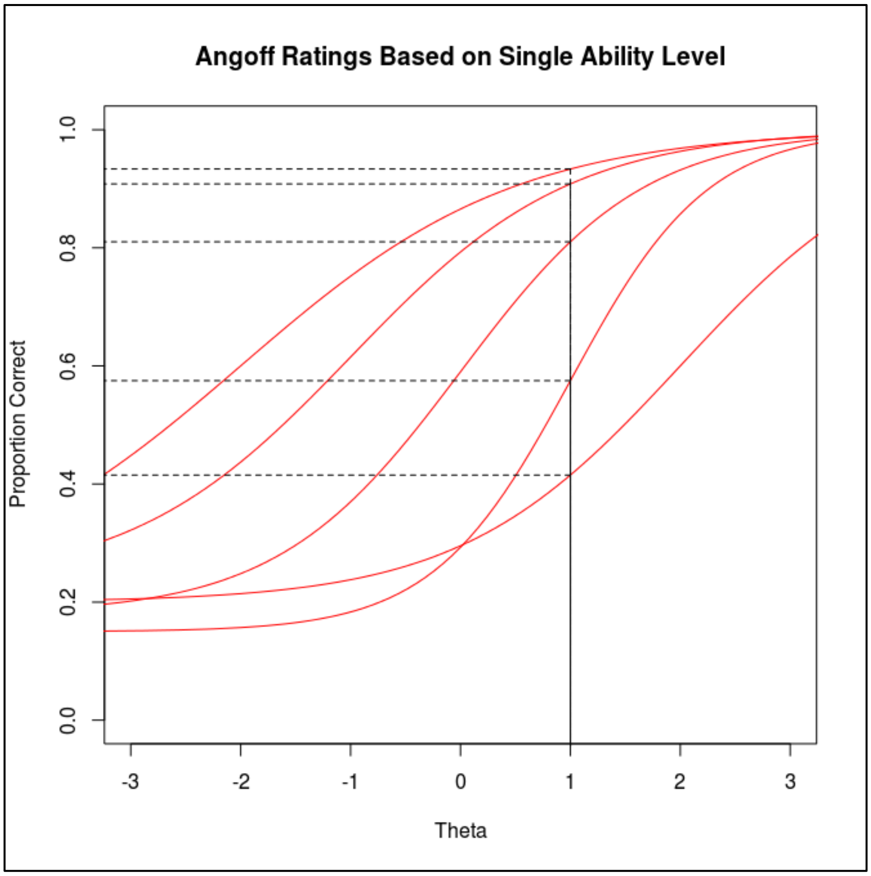

Figure 1 Expected Angoff Ratings for the Internally Consistent Judge

In Figure 1 the vertical line indicates one judge's view of the minimally acceptable examinee on the ability scale. The item characteristic curves show that although the judge has internalized a single ability level, the probability of a correct response on each item are influenced by the item parameters. A judge exhibiting perfect internal consistency would produce probability estimates at the intersection of the ICC with the examinee's ability.

The goal of any standard setting procedure is to place expert judgment on the score scale. In the case of the Angoff method applied in an IRT framework, the goal is to use a judge's ratings to estimate the proficiency score (i.e., "theta score") for the minimally acceptable examinee. To estimate this underlying theta score the probability estimate for each item is mapped through the ICC to arrive at a theta score. In Figure 2 a judge has provided ratings for five items each with different item parameters. Although the ratings do not result in a single theta estimate they indicate that the probable location of the judges internalized ability of the borderline candidate it approximately 1.0 on the IRT proficiency scale.

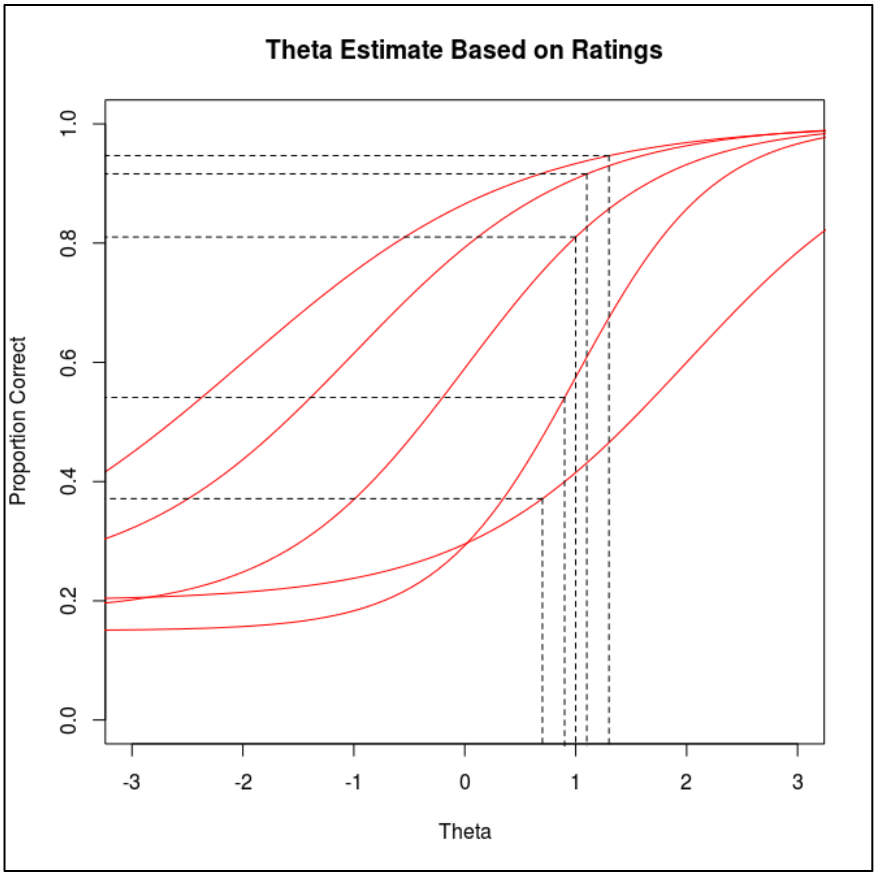

Figure2 Item Level Theta Estimates Bases on Angoff Ratings

These item level theta scores can be viewed as individual estimates of the judges internalized proficiency level. The estimate of the judge's internalized proficiency level is the median (or mean) of his or her individual item level theta estimates. The panels recommended cut score, is the median (or mean) item level estimate across all items and judges.

1.3.2 Properties of the IRT-Based Angoff Passing Scores

Stability of Performance Standards. In the typical Angoff method, research has consistently shown that judges struggle to produce internally consistent results. Some authors have viewed this problem as so significant that they consider it to be a fatal flaw of the Angoff method (e.g., Shepard, 1995). Although this limitation has some support in the measurement literature, also, the Angoff method does nothing to reflect the uncertainty in each judge's internalized performance standard. Judges' ratings are treated as conditional p-values for the minimally proficient examinee when measured without error. These values cannot be individually linked to student ability and instead must be considered in aggregate in the form of an expected test score. These expected test scores are averaged across judges to arrive at the panel's recommended cut score.

Ratings within the IRT Angoff framework, on the other hand, are not an element of an examinee's item or test score. Instead each individual Angoff rating can be mapped through the item characteristic curve to provide an estimate of the judge's internalized cut score on the IRT proficiency scale. Estimating the judges' internalized performance standard at the item, rather than test level, allows for the panel cut score to reflect the complete distribution of the judges rating rather than relying solely on each judge's imprecise point estimate. When developing the panel's recommended passing score, these individual distributions of judge's ratings can be combined into a single distribution of cut scores. Using the median (or mean) of this distribution as the recommended panel cut score reflects the certainty of judges' ratings to provide a more reasonable and theoretically more internally consistent estimate of the panel's judgment.

To illustrate the relative stability of passing scores set using Angoff method within the IRT framework it is helpful to imagine a distribution of judgments for four judges on the IRT proficiency scale. These four judges have different internalized cut scores and varying levels of internal consistency but their ratings can be combined into a single distribution. When a fifth judge is introduced into the panel, her influence on the panel's recommended cut score is a function of the stability of her ratings.

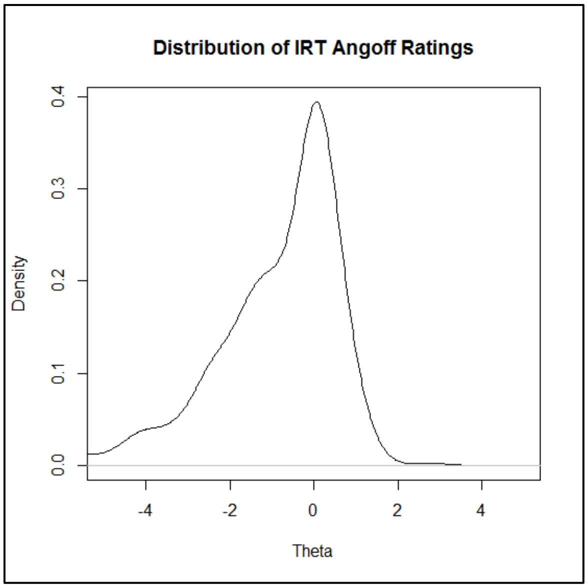

Figure 3 Add Inconsistent Judge

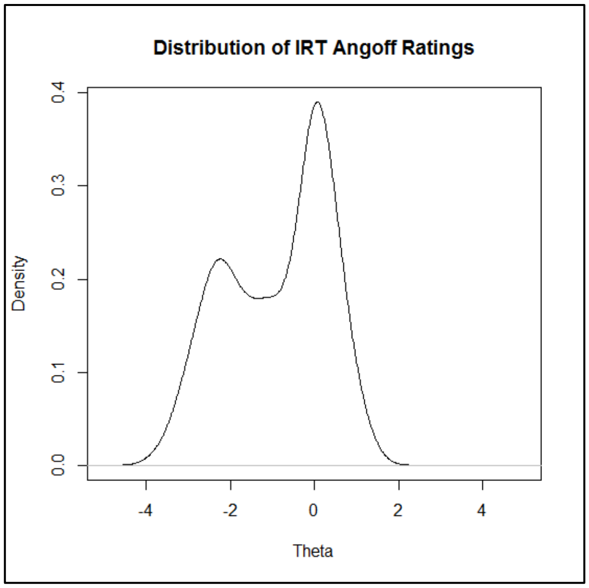

Figure 4 Add Consistent Judge

Table 1 Comparison of Mean and Median Passing Score

	Four Judges	Inconsistent Judge Added	Consistent Judge Added
Mean	-0.468	-0.843	-0.880
Median	-0.150	-0.411	-0.598

Figures 3 and 4 demonstrate the influence of this judge on the panel's recommended passing score. In both figures the additional judge has internalized a proficiency of -2.5 for the minimally capable examinee. The difference between Figures 3 and 4 is in the internal consistency of the judge's ratings. Figure 3 indicates the influence on an internally inconsistent judges on the distribution of ratings. Although this judge shifts the panel's recommended passing score to the left for both the median and mean, the magnitude of this change is significantly larger for the mean. In Figure 4 the inconsistent fifth judge has been replaced by a judge who produces highly internally consistent rating. In this case the median decreases significantly to reflect our certainty in the judge's opinion, but the mean remains virtually identical to the previous example. Because the median accounts for the spread of the judges' estimates it is differentially influenced by the consistency of ratings. By considering our certainty in a judge's view of the minimally competent examinee, the IRT-based Angoff approach could be used to increase the stability of the panel's recommended cut score.

Invariance. When using the common Angoff method each judge's assessment of what constitutes minimal proficiency is dependent on the set of reviewed test items. Although modern test theory treats ability as invariant to item selection, the common Angoff method on the test score scale fails to properly reflect this perspective. One common modification designed to address this limitation has been the mapping of the average cut score across panelists on the test score scale onto the IRT proficiency scale using the test characteristic curve. Although this approach does place the cut score on the IRT scale, it does not necessarily result in a passing score which is invariant to the selection of test items. Since passing scores are typically applied to multiple forms across several years of testing, these scores must be invariant to item selection. If passing scores are systematically influenced by item difficulty, the final passing score will fail to reflect the judges' expert opinion.

To develop item invariant passing scores, the IRT Angoff approach assumes that the judges' view of the minimally proficient examinee can be represented on the IRT theta metric. From this perspective the recommended cut score is not the theta associated with the judges' average test score, but instead is the median (or mean) of the judges' individual theta scores. Since theta scores directly drive the location of the recommended cut score, performance standards will theoretically be consistent across items. This invariance allows consistent passing scores to be set regardless of the specific set of test items.

Selective and Adaptive Standard Setting. True Score Angoff standard setting requires that all content experts provide ratings for each item regardless of their familiarity with the content or comfort with the task. This requirement is not only inefficient but has the potential to artificially bias the recommended passing score. With IRT, the Angoff cut score is estimated at the item rather than test level. This means that rather than relying on the total test score to estimate the ability of the minimally acceptable examinee, an examinee's ability can be estimated based on each item. This item level measurement means a judge's internalized performance standard can be inferred based on a subset of the total test. This feature provides two main benefits for practitioners: selective and adaptive standard setting.

In a selective standard setting procedure judges are allowed to omit items which they feel uncomfortable rating. Judges can choose to omit items on the basis of specific content, or uncertainty about how the minimally acceptable examinee would perform on the item. These self-directed item omissions do not ensure that judges will provide internally consistent ratings but they do eliminate the imperative that judges rate items outside their expertise. Although the effect of self-directed item omission has not been previously studied, the logic this approach is in keeping with the Angoff method which demands that judges are experts in the tested content. By allowing for the selective omission of test items, the IRT Angoff method ensures that judges feel they are content experts for all items for which they provide ratings.

In addition, item level measurement makes it possible for each judge's recommended cut score to be continually revised throughout the rating process as in an adaptive testing environment. In this way the Angoff method applied within an IRT framework allows for adaptive standard setting which has the potential to provide many of the same benefits as traditional CAT administrations for students. The primary benefit of adaptive standard setting is a reduction in administration time by omitting items which fail to provide information in the area of the cut score. For example, item with asymptotic ICCs in the area of the cut could safely be omitted as these items do very little to aid in the estimation of a judge's internalized performance standard. By eliminating the need for judges to rate uninformative items, adaptive standard setting could result in more precise passing scores with reduced administration time.

1.4 Statement of Problem

Developing valid cut scores is an integral part of the test development process for any examination used to make classification decisions. To establish cut scores subject matter experts decide what level of content mastery should be considered minimally acceptable. Standard setting is a systematic procedure for inferring these expert opinions and placing those opinions at an appropriate point along the score scale. Although many standard setting procedures exist, perhaps the most widely employed and studied method is the common Angoff method where the resulting cutscore is reported on the test score scale (Meara, Hambleton, & Sireci, 2001). Rather than asking judges about the importance of an item, or the appeal of specific response options, the Angoff method focuses on the minimally acceptable examinee's expected performance on each item. Although the intuitive appeal of the Angoff method is undeniable, concerns persist as to whether passing scores established with this method properly reflect the opinion of the content expert (Shepard, 1995; National Research Council, 1999).

Although the mechanics of the Angoff method are quite straightforward, the feasibility of the method is threatened by the inability of content experts to make the required judgments. Since item ratings are the mechanism through which the judge's expert opinion is inferred, inconsistencies in these ratings obfuscate the judge's true opinion. Specifically, when judges fail to produce internally consistent estimates of examinee performance, the individual ratings do not point to a single unique performance standard. Instead, these ratings may indicate examinees of dramatically different abilities are all minimally acceptable. Unfortunately these estimation errors are ignored during the calculation of the final recommended passing score. The impact of these errors is a function of the nature of the errors. When errors are random and symmetric, the final passing score may be quite reasonable. Alternatively when these errors are skewed, the common version of the Angoff method will produce a bias estimate of the judge's expert opinion.

In addition to the difficulty of the rating task, the item dependent nature of the common Angoff method can artificially influence the placement of the final passing score. Although the judges' belief about the ability of the minimally acceptable examinee is theoretically independent of item difficulty, the specific scale transformation used in the common Angoff method fails to ensure score invariance across items. Instead judges with consistent views of examinee ability could develop meaningfully different performance standards solely as a result of item selection. Passing scores developed in this fashion will not properly reflect the opinions of the content experts, since item dependence has the potential to significantly influence the position of the resultant passing score.

Finally, the commonly applied Angoff method is limited by its requirement that judges provide ratings for all items. This requirement is not only inefficient but has the potential bias the estimates of the judges view of the minimally acceptable examinee. When judges are forced to provide performance estimates for items they do not feel qualified to rate, the ratings will fail to properly reflect the judge's informed opinion. Furthermore it is reasonable to expect that these errors will tend to compound across items since judges can be expected to inflate the difficulty of items outside their area of expertise. Although the magnitude of this problem will depend on the composition of the panel and the tested content area, the requirement that judge rate all items has the potential to meaningfully bias the final passing score.

The goal of standard setting is to infer the opinion of content experts and to reflect that opinion as a point along the score scale. Unfortunately the common Angoff method has several limitations which interfere with its ability to properly estimate the judge's expert opinion. These limitations are brought on by lack of rating consistency, score invariance, and item level measurement. Each of these limitations has the potential to bias the estimate of the judge's view of the minimally acceptable examinee. Despite the centrality of test passing scores to the valid interpretation of test scores, these limitations suggest that passing scores established using the common Angoff method may fail to reflect the informed opinions of the panel of content experts.

1.5 Statement of the Purpose

The purpose of this study is to examine the benefits of interpreting Angoff ratings within an item response theory (IRT) framework. Although IRT has been used in the past to evaluate Angoff ratings, this would represent the first comprehensive analysis of the measurement properties of Angoff passing scores set within a modern test theory framework. Although theoretically, interpreting Angoff results using item response theory, has the potential to mitigate many of the limitations of the commonly applied Angoff method, these benefits have not been demonstrated in practice. This study will compare the Angoff standard setting results across two frameworks, classical (or test score) and IRT, to determine if the IRT based performance standards result in greater stability, flexibility, and efficiency. Successful completion of this study could have important implications for how passing scores are set and evaluated. When standardized tests are used to make high stakes decisions, the outcomes from this research could lead to more accurate decision making through setting more valid and reliable passing scores.

1.6 Outline of the Study

This dissertation contains five additional chapters. In Chapter Two the relevant literature on the Angoff standard setting method will be reviewed with specific attention devoted to the limitations of this method, and potential benefits of an IRT-based standard setting approach. Chapters Three, Four, and Five will present the methodology and results for three studies designed to assess the potential benefits of using the IRT Angoff method. Collectively these three studies will provide an examination of the advantages and disadvantages of interpreting Angoff ratings within an item response theory framework. The final chapter will provide a discussion of the results from these three studies as well as overall discussion and recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Literature Review

The Angoff method is an iterative test-centric procedure for estimating the content experts' recommended performance standard. Although the method as first described by William Angoff of Educational Testing Service in 1971 (Angoff, 1971) has been subject to a wide variety of modifications the method as commonly employed today includes four primary phases. Judges begin by discussing and internalizing the proficiency of the minimally competent examinee (MCE). In the second phase judges estimate the performance of the MCE for each test item. Next, these estimates are revised with the support of group discussion and typically empirical data of some kind. Finally, the individual item ratings are combined across judges and translated to a point along the score scale. This chapter begins with a review of the literature on each of these four phases with specific attention devoted to how each of these phases support the overall validity argument for the resulting passing score. Finally, an examination of two modern adaptations to the Angoff method, IRT estimation of performance standards and dynamic item selection procedures, will be addressed.

Overall the chapter will be divided into six sections:

1. Conceptualizing the MCE. This section will discuss the judges' ability to internalize the ability of the minimally competent examinee. Its focus will be on a number of studies presenting survey results for judges throughout the judgment process.

2. Internal Consistency of Judges' Ratings. This section will discuss the Angoff item rating process. Its focus will be on evaluating the validity of the passing score by examining the internal consistency of judge's ratings.

3. Feedback Between Rounds. This section will examine how the provision of empirical examinee performance data between rounds impacts passing scores. These paragraphs will devote considerable attention to judge's ability to integrate empirical data without devolving into norm-referenced judgments.

4. Placing Angoff Ratings onto the Test Score Scale. This section will briefly discuss some of the techniques for translating test score performance standards onto the IRT proficiency scale. It will devote specific attention to the potential for item dependent standards when examinee true scores are placed on the IRT proficiency scale.

5. Setting Angoff Standards Using IRT. This section reviews the literature on the use of IRT for setting Angoff passing scores. This will include a discussion of how IRT has been used to inform judgment weighting procedures as well as how IRT allows for the direct estimation of a judge's recommended performance standard on the IRT proficiency scale. In addition this section will illustrate how the IRT Angoff approach presented in this manuscript compliments and extends earlier discussions on this topic.

6. Selective and Adaptive Rating of Test Items. This section will examine the practicality of setting passing scores on a subset of test items. This will include specific discussion of selective standard setting in which judges may skip items and adaptive standard setting it which judges rate items selected algorithmically.

2.2 Conceptualizing the Minimally Capable Examinee

In test-centric standard setting methods judges begin by determining the level of content mastery which should be considered minimally acceptable (Livingston & Zieky, 1982). This process typically includes a detailed discussion of the knowledge, skills, and abilities which would be exhibited by the borderline examinee. Furthermore, these discussions may be supported by predefined performance level descriptors which broadly outline the ability of examinees in each performance category. The goal of these discussions is for panelists to determine and internalize a single ability level which they deem to be minimally acceptable. The estimation of this implicit ability is extremely important since it ultimately yields the explicit point on the score scale which will serve as the passing score.

During this process the panel of judges may not arrive at a single view of the minimally capable examinee. Although judges are typically selected on the basis of their familiarity with the examinee population and test content, at times other political and practical concerns influence the composition of standard setting panels (Reckase, 1998). Even when panelists are unquestionably content experts, variations in their professional experience may lead them to legitimately different views of minimal proficiency. Many researchers have recommended that facilitators encourage consensus within the panel (Livingston & Zieky, 1982; Plake, Melican, & Mills, 1991). Although this may be attractive from a measurement prospective at times, a single view of the MCE may be unrealistic. This disagreement across judges does not pose an inherent problem, provided that the panel of content experts can reasonably be viewed as a random sample from a pool of potential judges (Clauser, Clauser & Hambleton, 2012; Davey, Fan, Reckase, 1996). Although these differing views have relatively little effect on the ultimate passing score, understanding how content experts conceptualize the MCE it an important element in evaluating the results of the standard setting method.

Skorupski and Hambleton (2005) were the first to explore how judges' view of the MCE evolves during the standard setting exercise. Data for the study were collected through a five part survey administered at key points in the standard setting meeting for an elementary ESL examination. Judges' impressions of the minimally acceptable examinee were evaluated both before instructions and after discussion of the MCE. The results indicate that prior to instruction judge's had dramatically different views of what behaviors would be associated with each of the four performance categories. For example, some judges believed that a Basic examinee had no English ability at all, while others believed he or she would be able to speak simple sentences. Even after further training and discussion of the performance level descriptors this discordance between judges persisted. Although views of reading ability began to coalesce across judges, the descriptions of writing performance at each level continued to represent a wide range of proficiency. These results would seem to support the notion that content experts do not share a single view of proficiency for the minimally capable examinee. Although training does help to harmonize judges' opinions, preconceived notions of ability can continue to influence judges ratings.

Although Skorupski and Hambleton (2005) studied how judges' views of the MCE evolve through training, the authors did not examine the consistency of judges during the ratings process. Giraud, Impara, and Plake (2005), on the other hand, compared judges' descriptions of the MCE before and after the operational standard setting task. The goal was to see how content experts' opinions of the MCE changed during the standard setting process. The authors describe two parallel studies, in reading and math, in which panelists collectively described a domain of skills associated with the MCE. These skills were recorded but were not available to judges as they provided their ratings. During the rating process judges were reminded to imagine how a single MCE would perform on each item. After all judgments were collected, panelists were asked individually to describe their internalized examinee. Results of this study indicate that judges' descriptions of the MCE closely aligned with the skills described originally. This was particularly true for the mathematics examination which had a detailed performance level descriptor for the proficient examinee. These findings suggest that the training process helps content experts to internalize a view of the MCE which could be maintained throughout the ratings process.

Although the authors acknowledge that more research is needed, these results imply that judges are capable of internalizing a single examinee ability and applying that ability throughout the ratings process.

2.2.1 Summary

Although relatively little empirical research has been conducted on judges' ability to internalize the MCE, the results of these studies offer some interesting conclusions. Skorupski and Hambleton (2005) found that not only can content experts enter with disparate views of the MCE, but at times they will fail to reach consensus even after training. Giraud, Impara, and Plake (2005) found that after internalizing the knowledge, skills, and abilities of the MCE, judges applied that set of abilities to the estimation of examinee performance. Together these studies suggest that although content experts are capable of imagining the MCE, they will not always agree on a single examinee ability. This consistent view of examinees' ability, however, does not imply that judges are able to produce ratings consistent with a single ability. Even when judges clearly imagine a single examinee's ability, errors in the rating process could result from difficulty in estimating that examinees performance. A discussion of the internal consistency of judges' Angoff ratings is considered next.

2.3 Internal Consistency of Judges' Ratings

After content experts have internalized the ability of the MCE, the judges review each test item and provide an estimate of how borderline examinees would perform. These estimates can be binary (correct/ incorrect), but more commonly they are estimates of the proportion of minimally qualified examinees who would answer the item correctly. For polytomous items judges typically provide estimates of the average score which would be achieved by the borderline examinees. Regardless of the particular method used, the sum of these estimates equals the borderline examinee's expected test score on the assessment. Although the logic undergirding the Angoff method is straightforward and reasonable, actual implementation of the method has been criticized for its cognitive complexity (National Research Council, 1999). Many authors have suggested that content experts struggle to identify the performance of the MCE, and therefore fail to produce grounded ratings (e.g., Busch & Jaeger, 1990; Shepard, 1995 ; Clauser, Harik, et al., 2009; Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Clauser, Mee, & Margolis, 2011; Clauser, Swanson, & Harik, 2002).

Although there is no absolute criterion to evaluate the accuracy of judgments, many authors have considered the internal consistency of judges' ratings as an important framework for evaluating Angoff judgments (Kane, 2001; van der Linden, 1982; Goodwin, 1999; Plake, Impara, & Irwin, 1999; Smith & Smith, 1988). Internal consistency, in this context, is the ability of a content expert to provide probabilities of success which could reasonably belong to a single examinee. A judge fails to produce internally consistent ratings when, for example, she indicates that the MCE will struggle with empirically easy items, but will succeed on empirically difficult ones. van der Linden (1982) believed that lack of internal consistency resulted in capricious and indefensible cut scores (p. 296). Kane (1994) reaffirmed this belief by saying "[internally inconsistent results] do not provide a solid basis for drawing any conclusions" (p. 445).

In addition to Kane and van der Linden several other authors have discussed the importance of internal consistency in evaluating the results of an Angoff standard setting exercise (e.g., Goodwin, 1999; Plake, Impara, & Irwin, 1999; Smith & Smith, 1988, Clauser, Clauser, & Hambleton, 2012). This work has primarily been focused on the comparison on Angoff ratings to the empirical item difficulties as a measure of internal consistency. At times Angoff ratings have been compared to classical conditional p-values (Smith & Smith, 1998). At times this approach has been implemented by identifying examinees within some more-or-less arbitrary score band around the cut score (Plake & Impara, 2001). Other researchers have compared judges' ratings to conditional probabilities of success based on IRT item parameters (Clauser, Clauser, & Hambleton, 2012). Finally, at least one study directly compared judges estimated cut scores to the actual performance of students previously identified as minimally proficient (Impara & Plake, 1998). Regardless of the specific methodology employed results of these studies have consistently suggested that judges struggle to provide accurate estimates of borderline examinee performance. The following paragraphs in this section summarize the relevant literature on the internal consistency of judges' ratings.

Impara and Plake (1998) studied the ability of content experts to accurately estimate examinee performance in an Angoff standard setting environment. The study included 26 middle school science teachers who were asked to predict examinee performance on an end-of-course science exam for both typical and borderline passing students. The results of the study showed that teachers struggled to accurately predict student performance for both groups of students. For typical students, the teachers overestimated performance by more than three test score points: from a true performance of roughly 32.5 to a predicted performance of just over 36.0. For the students defined as "borderline passing," based on course grades, the teachers underestimated performance by nearly 10 points: from a true performance of approximately 22.5 to a predicted performance of just over 13.0. It should be noted that teachers involved in this study were intimately familiar with the course content, the end-of-course exam, and even the specific examinees about whom judgments were provided. These results suggest that even in what must be considered a best case scenario, judges struggled to make the required judgments.

Like Impara and Plake (1998), Plake and Impara (2001) compared judges Angoff ratings to actual examinee performance. For this study Angoff ratings from a financial management certification examination were compared to the true performance of examinees near the recommended cut score. For this sort of comparison it is not reasonable to look at the mean difference between judge's ratings and examinee performance, due to the inherent dependency which results from using the performance standard to evaluate the performance standard. It is possible, however, to examine the absolute difference between the Angoff judgments and true examinee performance. The results of this comparison showed that across both years of testing the mean absolute difference was 7%. Based on these results the authors concluded that a difference of this magnitude, represent a "very high degree of congruence" (94) between estimated and actual examinee performance. As a practical matter it is not clear what impact a difference of this size would have on passing rates. If all errors were in the same direction a 7 item difference on a 100 item test would seem to be a rather sizable error. If errors were more or less symmetric the impact on passing rates would be negligible, but this study does not provide a reasonable basis for judging the distribution of errors in judges' ratings.

Although the studies above, compare the judges' average ratings to examinee performance, they provide little insight into the ability of individual judges to estimate examinee performance. Clauser, Clauser, and Hambleton (2012), were the first to empirically examine the ability of individual judges to provide internally consistent estimates of examinee performance. The authors calculated the correlation between the judge's ratings and the model implied empirical conditional probabilities of success. Although other authors have used IRT based empirical conditional probabilities to evaluate Angoff ratings, this study was the first to calculate a unique set of probabilities for each judge. Therefore rather than comparing the group's average rating to the expected rating, each judge was compared to him or herself. The results of this study suggested that overwhelmingly, judges struggled to produce reasonable estimates of borderline candidate performance. Across two tests with three panels each, the correlation between judges actual ratings and the empirical probabilities never exceeded 0.60. Even more telling, roughly half of the judges produced ratings which had correlations which were not statistically different from zero. Together these results suggested that although there is a considerable range in judges' abilities to produce accurate estimates of examinee performance, many judges produced ratings which were essentially unrelated to the actual difficulties of the items. Of course the generalizability of the Clauser et al. results found in a medical exam context is not known.

2.3.1 Summary

The literature presented in this section suggests that concern over judges' ability to estimate examinee performance on particular items may be well founded. Although no absolute criterion to judge the accuracy of Angoff ratings exists, many authors have suggested that content experts be judged on the basis of the internal consistency of their ratings. Comparisons of judges average Angoff rating to conditional p-values have suggested that errors range from modest to quite substantial. When judge's ratings are considered individually rather than in aggregate, results indicate that many judges produce ratings which bare virtually no relationship to actual borderline examinee performance.

2.4. Feedback Between Rounds

Although the Angoff method only requires a single round of judgments, one common modification allows judges the opportunity to revise their estimates. In this modification to the traditional method the judgment process is divided into two or three rounds. Between rounds judges are often provided with examinee performance data such as p-values or performance deciles to help ground and inform their judgments (Clauser, Sireci, & Clauser, 2010; Hambleton & Pitoniak, 2006; Reckase,2001). Although this iterative procedure was not originally describe by Angoff, empirical results and practical experience suggest that judges feel more confident in the process when they are allowed to provide revisions. Furthermore several studies which have compared the internal consistency of Angoff ratings made before and after judges review performance data have shown that ratings show an increased correspondence to actual item difficulties (Swanson, Dillon, & Ross, 1990; Busch & Jaeger, 1990; Clauser, Swanson, & Harik, 2002). The following paragraphs will discuss the relevant literature on the influence of performance data.

Busch and Jaeger (1990) studied the effect of performance data on judges' internal consistency by evaluating the changes in the covariation between judges' ratings and item p-values on seven different tests. For each test, content experts were asked to judge each item first without examinee performance data and then after having an opportunity to compare their judgments to empirical item difficulties. The authors found that without performance data the judges' ratings correlated only modestly with p-values, averaging 0.60 across the seven tests. When performance data were introduced, however, correlations significantly increased across all seven tests, averaging 0.84.

Both Clauser, Swanson and Harik (2002) and Clauser, Harik, et al. (2009) mimicked Busch and Jaeger (1990). Both studies had judges review items without and then with data and compared the consistency of judges ratings. Unlike Busch and Jaeger (1990) these studies used IRT derived empirical conditional probabilities rather than p-values to assess judges' internal consistency. The two studies yielded similar results. Clauser, Swanson and Harik (2002) showed that without performance data the correlations between judges' estimates and conditional p-values were 0.63 and grew to 0.98 after performance data were provided. The results from Clauser, Harik, et al. (2009) indicate that without the aid of examinee performance data the correlation between judges' estimates and conditional probabilities was approximately 0.34. When performance data were introduced the correlation increased dramatically to 0.66. These results support the previous finding that judgments made without performance data had only a modest correspondence with empirical conditional item difficulties; after data were provided, judges' internal consistency increased substantially.

These studies seem to suggest that provision of examinee performance data has a profoundly positive impact on the consistency of judges' ratings. Some researchers, however, have expressed concerns that judges may rely too heavily on examinee performance data. This issue could be so severe that at times it may results in diluting the criterion-referenced performance standard with a partially or completely norm-referenced performance standard. Maurer and Alexander (1992) were among the first to express concern about the effect of performance data on the estimation of cut scores. In their study of the Angoff method, they evaluated several modifications to the traditional method, including the provision of performance data. Although the authors conceded that judges often exhibit low internal consistency, they argued that use of performance data may undermine the defensibility of the resulting passing scores. As the authors stated, "[t]here would seem to be a potential danger of judges abandoning their expertise in favor of using the normative data to generate judgments," (p. 774).

Two recent empirical studies have attempted to determine how judges make use of examinee performance data (Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Mee, Clauser, and Margolis, 2011). Ideally content experts arrive at judgments by leveraging their own professional experience to estimate the performance of the MCE on the specific test item.

When performance data are introduced, they then are expected to integrate these data into those content-based judgments. If judges mechanically bring their ratings in line with the data rather than integrating the data into their professional judgments, the final result may be little more than a norm-referenced performence standard. If this is the case it is difficult to defend the resultant passing score on the grounds that it is content based.

Clauser, Mee, Baldwin, Margolis, and Dillon (2009) conducted a study to try to better understand how judges actually use performance data. At issue was whether judges integrated performance data into their content-based judgments or deferred to the data to generate essentially norm-referenced performance standards. To test the manner in which judges utilized performance data, the authors asked judges to rate 75 items in two rounds. In round one, judges were asked to rate the items without the aid of performance data; in round two, performance data were provided and judges were given an opportunity to revise their estimates. What made this study unique was that for half of the items the performance data had been randomly shuffled from one item to another. If performance data truly served only to help the judges spot some new insight or understand some nuance of item performance, the manipulated performance data should have had virtually no effect on the judges' ratings. The results indicated that judges tended to alter items with manipulated and non-manipulated data to approximately the same degree. Overall, the authors concluded that judges either incorporate performance data mechanically with no ability to explain the results or build a personal rationale to explain away any perceived logical inconsistency.

In a follow up to Clauser, Mee, et al. (2009), Mee, Clauser, and Margolis (2011) designed a study to investigate whether the earlier results would have been different if the panelists had been given different instructions. The study followed the same methodology as that of the Clauser, Mee, et al. (2009) study with one important difference: rather than giving the judges typical instructions with regard to the performance data, the judges were cautioned that some of the data had been manipulated. The goal of the study was to determine if judges would make differential use of the manipulated and non-manipulated data when they knew to expect that some data may not be genuine. The logic was that telling the panelists that some of the data were inaccurate represented the strongest possible set of instructions they could receive. In this scenario, mechanically moving toward a norm-referenced judgment would seem to be irrational. The results indicated that, although the magnitude of the revisions tended to be smaller than those reported in the parallel study by Clauser, Mee, et al. (2009) the relationship was comparable for items presented with manipulated and non-manipulated data.

2.4.1 Summary

This section considers the impact of one of the common Angoff modification in which judgments can be revised across multiple rounds. Between these rounds judges are typically provided with examinee performance data and given an opportunity to revise their ratings. Several studies have shown that this type of feedback has a profoundly positive impact on the internal consistency of judges rating. However concerns persist as to whether Angoff ratings made in the presence of prescriptive performance data can be considered criterion-referenced. Recent studies have suggested that judges may rely too heavily on performance data, rather than relying on their own professional judgment. Insofar as these results are broadly generalizable it suggests that provision of performance data may result in a passing score which is at least partially norm-referenced.

2.5. Placing Angoff Ratings onto the Score Scale

For most modern certification and licensure examinations, after all Angoff ratings have been collected, and the forms have been equated, the test score performance standard must be translated onto the IRT proficiency scale. This transformation allows for passing scores which can be consistently applied across multiple test forms. By far the most common method for this test score to scaled score translation is simply mapping the recommended test score through the test characteristic curve to a point on the IRT proficiency scale (Reckase, 1998; Davey, Fen & Reckase, 1996). Although this approach is straightforward and has some intuitive appeal, it may be limited due to the item dependent nature of the resulting passing score.

The primary role of the test score to scaled score translation is to ensure that passing scores are applied consistently regardless of which items appear on a particular form of the test. Mapping solutions which do not produce item invariant results are therefore suspect. Although this point has not been thoroughly discussed in the literature, Ferdous and Plake (2008) pointed out that the theta associated with the average item rating is not necessarily the same as the average of the individual theta estimates for that item. Although the grand mean of Angoff ratings is unchanged, the authors explain that “due to the nonlinear relationship between item performance estimates and IRT ability estimates, these methods may not yield identical results” (Ferdous & Plake, 2008, pp.781). To illustrate the item dependent nature of the resulting scaled score performance standards it is helpful to imagine judges providing Angoff ratings for two different one-item tests.

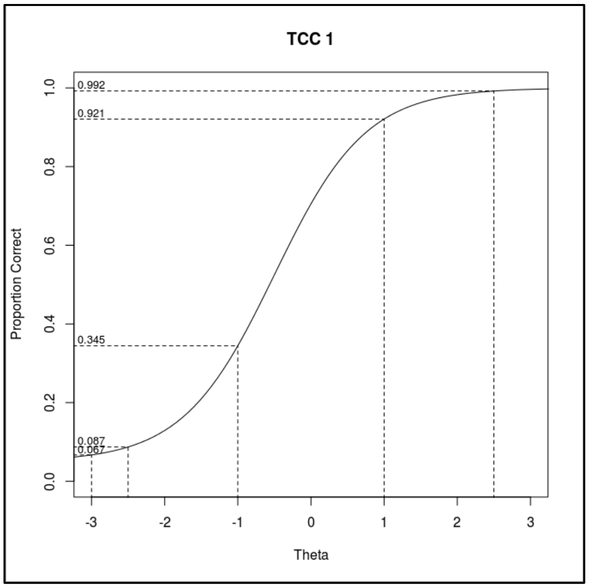

Figure 5 Test Characteristic Curve 1

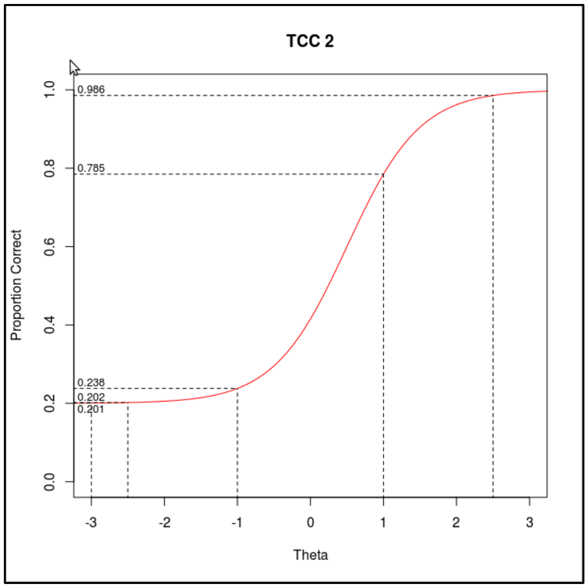

Figure 6 Test Characteristic Curve 2

In the above figures (Figures 5 and 6) five judges have internalized different passing scores which can be represented as five distinct points on the IRT proficiency scale. Although each judge's belief about the ability of the minimally proficient examinee is unaltered across items, differences in the item parameters result in different Angoff ratings. These ratings are averaged to arrive at the passing score on the test score scale, which is then mapped onto the IRT proficiency scale through the test characteristic curve. In this example, despite the differences in individual judges' ratings, the average of the judges ratings on the test score scale are equal across both items (0.482).

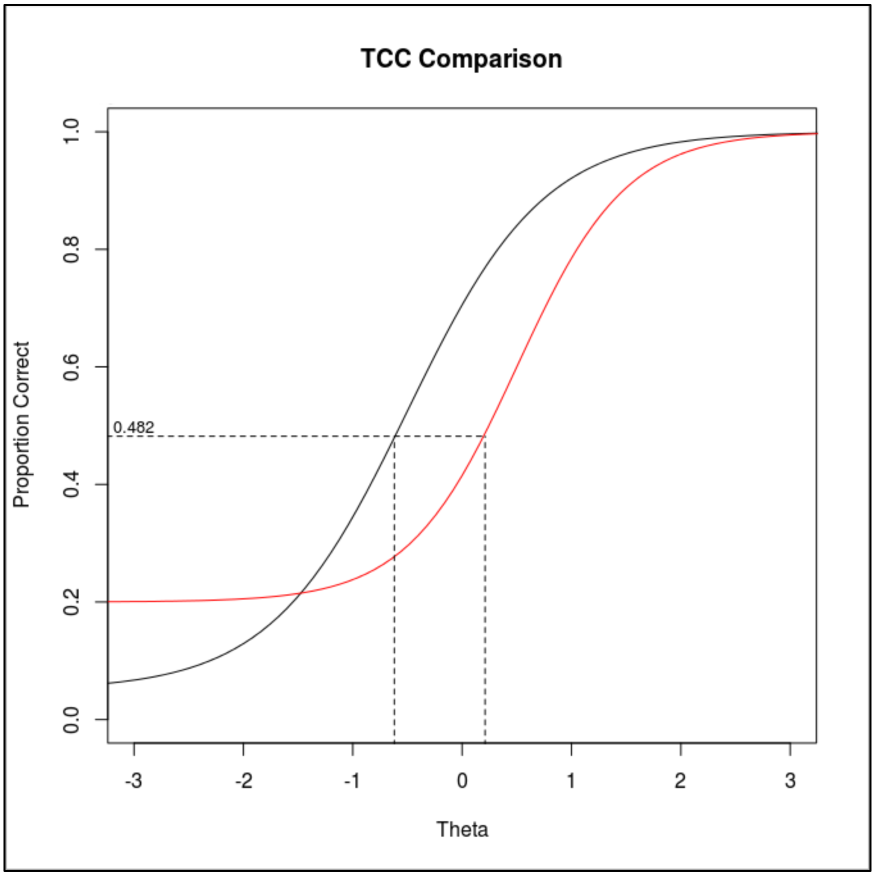

Figure 7 TCC Comparison

When this average test score is brought back onto the IRT proficiency scale through the TCC the two items result in dramatically different theta values, despite the fact that the judges' view of the minimally acceptable examinee had remained consistent. The magnitude of this difference will be affected both by the linearity of the TCC and the spread of the judges opinions. Although this example is an extreme case, it clearly highlights the potential impact of test items on the final recommended passing score.

The distortion of the judges ratings which results from this test score averaging followed by a non-linear translation onto the IRT proficiency metric potentially undermines the credibility of the resulting passing score. Reckase (2000) points out that one technique for evaluating the standard setting method is in the degree to which it preserves the content experts' judgments. As Reckase notes "The question is whether the standard-setting method can recover the theoretical cut-score assuming a judge performed every task consistently and without error" (50-51). The above example illustrates a potential limitation of this score translation procedure, as judges providing consistent ratings could readily arrive at different passing scores based solely on the reviewed items. Although equating takes place to ensure that all items are on the same underlying scale, this process does not result in identical TCCs for tests of different difficulty.

2.5.1 Summary

This section highlights a virtually unstudied potential source of instability in Angoff performance standards. In addition to errors in judges ability to estimate the performance of the MCE, peculiarities of the non-linear transformation of the raw score performance standard onto the score scale has the potential to further distort judges opinions. This feature means that passing scores established using the Angoff method on the test score scale will not be invariant across the items selected. Although these errors may be fairly minor their influence potentially undermines the validity of the passing score by distorting the judgment of the content experts.

2.6 Setting Angoff Standards Using IRT

Although the direct mapping of test scores through the TCC is by far the most common technique for translating test score performance standards onto the latent trait scale, several alternative methods have been described. These methods rely on item response theory to provide a weighted average of judges' ratings or estimate the passing score directly based on item level theta estimates. Although these approaches have not been commonly adopted, the use of IRT may offer considerable appeal to practitioners. These IRT-based techniques more properly reflect the scoring methodology employed in many testing programs. In addition, since the recommended passing score is determined in terms of theta estimates rather than on the test score, the resulting passing score should be truly independent of the reviewed items. The following paragraphs will be divided into two sections. They will begin by outlining how IRT has been used to weight Angoff ratings to arrive at a recommended panel cut score. This will be followed by brief discussion of how IRT has been used in the direct estimation of the scaled score performance standard based on item level theta estimates.

2.6.1 IRT Weighting

As typically implemented, the Angoff method assumes equal weights for all judges and items. Although this approach is well established and reasonable, several authors have questioned its defensibility when faced with diverse sets of judges and items (Davey, Fen, & Reckase, 1996; Clauser, Clasuer, & Hambleton, 2012). These authors have pointed out that since judges vary considerably in their internal consistency and items vary in the accuracy of their parameter calibration, it may not be appropriate to simply average all ratings. With this in mind, several researchers have explored the possibility of providing a weighted average of individual ratings to arrive at the panel's recommended passing score (Kane, 1987; Plake & Kane, 1991; Davey, Fen, and Reckase, 1996; Skorupski, 2012).

Kane (1987) was one of the first researchers to consider the application of item response theory to judgmental standard setting methods. Kane argued, provided that the IRT model fit both the student response data and the Angoff judgments, the ability of the MCE could reasonably be considered a point on the IRT proficiency scale. With this in mind Kane presented two techniques for using Angoff ratings to estimate a single recommended cut score on the IRT proficiency scale. Method one determined the cut score on the proficiency scale to be the point that minimizes the squared differences between the empirical probability of a correct response, and average ratings across judges for each item. This method implicitly provides weights based on the consistency in judges ratings for each item. Method two also averaged Angoff ratings across items, but unlike method one, these probabilities were mapped directly onto the IRT proficiency scale through the item characteristic curve. These item level thetas were then weighted in proportion to the inverse sampling variance to arrive at the panel's recommended cut score. Each of these procedures results in a recommended passing score that is influenced by the consistency of item level ratings across judges.

To evaluate the efficacy of these models Plake and Kane (1991) examined Angoff results using a simulation study. In this procedure that authors simulated Angoff style responses using the three-parameter model with the addition of both systematic and random error. The results of the simulation study indicated that neither of these weighting techniques meaningfully improved the estimates of the known true passing score over the more traditional unweighted approach. The authors conclude that direct mapping of test scores through the TCC is the most appropriate method in practice given its relatively simple implementation.

Despite the results of Plake and Kane (1991), interest remains in Kane's earlier work. Davey, Fen, and Reckase (1996) evaluated the stability of Angoff performance standards developed using a variation on Kane's least squares technique. This approach determined the passing score by finding the point on the IRT proficiency scale which minimized the differences between judges' ratings and the empirical probabilities. Unlike Kane (1987) however, the authors performed a logit transformation on both Angoff ratings and empirical probabilities. This transformation was found to produce a nearly normal distribution across ratings and equalize variances across items. It therefore was not necessary to control for sampling variance in the estimation of the passing score. The authors used a jackknife simulation in which individual judges were dropped from the estimation of the passing score. Although the accuracy of the passing score cannot be evaluated, the results of the simulation study showed that the least squares procedure produced considerably more stable cut score estimates than the TCC mapping approach.

Clauser, Clauser, and Hambleton (2012) explored the possibility of weighting Angoff judgments based on each judge's internal consistency. The authors compared each judge's item ratings to the IRT defined empirical probabilities. Each judge's recommended cut scores was then weighted proportionally to reflect the judges internal consistency. The results of these analysis showed that across two different data sets the panel-level internal consistency increased. Perhaps more interesting, however, is that for both data sets the recommended cut scores across panels converged. Although the authors considered this result promising, they note that further research would be required to illustrate its generalizability.

2.6.2 Direct Theta Estimation

Although weighting Angoff ratings may be appropriate from a pure measurement perspective, practical considerations often undermine its defensibility. Implicit in any weighting scheme is the notion that judges and item are each being selected at random from a pool of qualified candidates (Clauser, Clauser, & Hambleton, 2012). Unfortunately, when items are selected to reflect content areas in specific proportions, or judges are selected to represent specific constituencies, this assumption is violated. Under these scenarios, a weighted passing score would fail to reflect the desired makeup of judges and items. Fortunately, several authors have considered methods for the direct interpretation on

Angoff ratings onto the IRT scale (Smith, 1999; Ferdous and Plake, 2008; Gross & Wright,1986).

Smith (1999) explored the possibility of comparing judges' "yes/no" Angoff-like ratings to IRT-based probabilistic response strings. In this study a probabilistic set of 0/1 item responses were generated for each item at 20 points along the IRT proficiency scale between -2 and 2. These response strings were then compared to judges yes/no Angoff ratings to identify the point on the IRT proficiency scale which best matched their responses. The results of the analysis showed that judges struggled to produce ratings which mirrored those produced by the IRT model. At times judge's ratings suggested that they could reasonably believe the cut score to exist anywhere within a 0.4 range on the proficiency scale. These results led the author to conclude that this is not a practical alternative to traditional classical approaches to Angoff estimation.

Despite the results of Smith (1999), 1/0 response data have been used effectively to estimate cut scores on the IRT proficiency scale. Ferdous and Plake (2008) considered the treatment of yes/no Angoff ratings as item responses. This approach allows for judges response pattern to be scored, to arrive at an estimated cut score on the proficiency scale. Although this method is quite appealing, it continues to rely on judges' ability to produce ratings consistent with a single ability level. Results of this study indicated that passing scores obtained using the response vector approach were very similar passing scores to those obtained using traditional Angoff methods. Across the two examinations studied, the variations in passing score would have effected passing decisions for at most 4% of examinees. Although the authors note that there is no empirical method to determine which of these passing scores are preferable, consistency across the scoring algorithm and standard setting methodology is desirable. Therefore for tests scored using IRT, the response vector approach may be appealing.

In addition to the response vector approach Ferdous and Plake (2008) was the first study to place each judge's individual item rating directly onto the IRT proficiency scale. This "judge theta" approach is appealing because it utilizes probability information, typically provided by Angoff judges, rather than utilizing only dichotomous ratings. In principle, individual Angoff ratings are placed into the IRT proficiency scale through the item characteristic curve. These ratings are then averaged across judges and items to arrive at a recommended passing score. As a practical matter rather than assigning the theta value which corresponded directly to the probability estimate, the authors opted to round each rating to the nearest of 61 quadrature points between -3 and 3 based on a normal distribution with a mean of 0.0 and a standard deviation of 1.0. For most items this approach is not likely to have a significant influence on the overall passing score but it does mitigate the effect of extreme values on the mean performance standard. Results of this study indicated that setting standards in this fashion yielded identical test score results to the classical Angoff approach. The authors suggest however that despite the consistent passing score this approach has the "potential for being the best match to the 3PL scoring model" (Ferdous & Plake, 2008, pp.785).

The judge theta approach described by Ferdous and Plake (2008) has several advantages and it bares considerable similarity to the IRT Angoff approach presented here. The primary difference between these methods is the selection of a measure of central tendency. Ferdous and Plake selected the mean as the method for summarizing the distribution of theta values. Although this is a desirable measure of central tendency for performing inferential statistics, the mean is not appropriate for the description of an asymmetric distribution. It has been well established that when dealing with an asymmetric distribution, the median is a more reasonable estimate of a typical value (Hays, 1994; Gravetter & Wallnau, 2009). Furthermore, when translating Angoff ratings onto the IRT proficiency scale there is a realistic possibility that there is no mean. When using the three-parameter model judges can readily produce estimates which will fall at +/- infinity on the IRT proficiency scale. For example if a judge estimated that the probability of a correct response is 0.18 for an item with a lower asymptote at 0.20 the corresponding theta value would be negative infinity, making calculation of a mean impossible. Although infinity could be replaced by an extreme real number, as was done by Ferdous and Plake (2008), the choice of value will often have a sizable effect on the resulting mean. The median's use of rank order, as opposed to absolute scores makes it capable of readily accommodating these extreme values. This suggests that the median theta may provide both theoretical and practical advantages over the judges' mean theta.

2.6.3 Summary

This section outlines the previous literature on the use of item response theory in setting test-centric performance standards. Although none of these approaches have seen widespread adoption, the use of IRT scoring algorithms strongly suggests that IRT should be used in the standard setting process too. Two broad approaches are described for the use of setting IRT Angoff performance standards: IRT weighting and direct estimation. The IRT weighting procedures allow the stability of judges or items to influence their effect on the passing score. Although this is promising, questions remain regarding the feasibility in many testing contexts. Direct estimation uses item responses on the IRT proficiency scale to estimate the overall passing score. These procedures have provided mixed results, but remain a potentially promising strategy for using IRT to set performance standards. In the next section, some of the potential benefits of IRT based standard setting methods will be considered.

2.7. Selective and Adaptive Rating of Test Items

One of the primary advantages of item response theory over true score theory is lack of item dependence in ability estimation. This item independence has significant implications for the development of performance standards using IRT. Simply put, IRT based standard setting methods should not require that all judges provide ratings to all items. The ability to set item independent cut scores has two obvious applications. First, unlike the typical test centric procedures where judges must provide ratings for every item, IRT would allow judges to select which items they wish to rate. This selective standard setting procedure would allow judges to forgo rating items which they felt unable to properly estimate, due to a lack of necessary expertise with the item content. Alternatively, item independent measurement would allow for judges to review items dynamically selected through a predefined algorithm. This adaptive standard setting method could provide a significant reduction in the time require for judges to review items, by adaptively selecting items which provide relevant information in the area of the judge's internalized performance standard.

The concept of judges responding to a subset of administered items is not a new one. Several authors have examined the consistency of standard setting result based on a subset of test content. Results of these studies have consistently shown that, provided the subset preserves the original tests difficulty and content coverage, the resulting passing scores are quite stable. For example Plake and Impara (2001), and Ferdous and Plake (2005) each looked at the consistency of passing scores developed on parallel split halves of the full length exam. Under these, relatively restrictive conditions, test score performance standards were found to be consistent across forms. In a more extreme case Sireci, Patelis, Rizavi, Dillingham, and Rodriguez (2000) compared the passing scores developed with a subset of the CAT item bank to those developed using the entire 120 item bank. Despite the fact that items were not selected to mirror the content of the full test, the authors found that passing scores set using only 80 items were within one-tenth of a standard deviation, of those set with the entire bank. Passing scores remained within two-tenths of a standard deviation of the full bank, with subsets as small as 40 items. The authors suggest that accurate passing scores could reasonably be set with even a smaller number of items, provided that items were selected intelligently to mirror the content and statistical characteristics of the complete bank. The following paragraphs will provide a brief review of the relevant literature on selective and adaptive standard setting.

2.7.1 Selective Standard Setting

Although selective standard setting has not specifically appeared in the literature, literature on related topics has implicitly called for it. With the Angoff method administered in the typical fashion, judges are obliged to provide a rating for each item regardless of their familiarity with the specific content. Often this may be as simple as not understanding the item well enough to provide an accurate estimate of examinee ability. At times however, judges may not know the correct answer to the item, and therefore would have little basis for providing a probability estimate. Although it is reasonable to expect that judges would provide internally inconsistent ratings for items outside their domain of mastery, these errors in judgment are not likely to be symmetric. Instead research by Saunders, Ryan, and Huynh (1981), has shown that judges conflate their personal lack of facility with the item and objective item difficulty. Specifically judges set lower passing scores for items they cannot answer and higher passing scores for items they can. Ryan, and Huynh found a correlation of 0.30 between judges achievement and their recommended passing score, which accounted for 9% of the observed variation in passing scores across judges.

Although these results make it clear that passing scores will vary based on the judges' level of content mastery, they do not provide empirical evidence for which passing score is correct. Theoretically arguments could be made for a passing score set using only items that the judge could answer correctly, or for a passing score set using all tested material. Chang, Dziuban, Haynes, and Olson (1996) thoroughly explored changes in both performance standards and the internal consistency of ratings across items the judges answered correctly and incorrectly. The results indicate that even after controlling for item difficulty, judges tend to produce higher passing scores for items they answer correctly than for those they did not. Furthermore the authors found that judges produced more internally consistent ratings for item they answered correctly. These results suggest that passing scores established using only item the judges answered correctly have the potential to be empirically more valid and defensible. Although more research is needed, these results suggest that a selective standard setting method in which judges could skip items which fell outside their area of expertise, may improve the internal consistency of performance standards by removing random or systematic errors with no additional burden on judges.

2.7.2 Adaptive Standard Setting

Unlike selective standard setting, in which judges respond to items of their choosing, adaptive standard setting algorithmically selects items for review. Although virtually nothing has been published on adaptive standard setting, these issues has been briefly addressed in Sireci and Clauser's (2001) exploration of a method for setting performance standards on computerized adaptive tests (CAT). Adaptive tests select test items dynamically during the testing process to minimize the standard error of measurement.

Although this approach is extremely powerful, performance standards set on the test score scale using traditional standard setting methods are inappropriate for an adaptive exam since test forms differ systematically in difficulty. Furthermore, for testing programs with large item banks, asking judges to provide ratings for each item would be impractical. To address these limitations Sireci and Clauser present a method suggested by Howard Wainer in a personal communication. In this Wainer Method, judges rate items in a completely adaptive environment, by providing dichotomous estimates as to whether the minimally competent examinee will answer the item correctly. These predictions are used with a traditional CAT routing algorithm to provide items which most closely mirror the estimated ability of the borderline examinee. This is the first and only discussion in the literature of a truly adaptive standard setting method. Although this technique sounds promising, no empirical research has examined its feasibility. Furthermore, the authors note that by asking judges to produce only dichotomous estimates of examinee ability some information may be lost.

Before any adaptive standard setting method can be adopted, it is important to demonstrate that the order in which the items are presented does not have a meaningful impact on judges' ratings. Plake, Impara, and Irwin (2000) examined these issues in their exploration of judge consistency across years. In that study a group of judges were impaneled in consecutive years to set performance standards for the same exam. Although the test forms had changed across years, a selection of year one items were embedded in the year two test for comparison purposes. The judges found that even with the elapsed year, changes in the order of test items had a trivial effect on judges' ratings. Specifically the authors found a mean absolute difference of 0.05 (in the p values) between year one and year two ratings. These results suggest that within a single year the order of item presentation is likely to have a negligible impact on judges Angoff estimates.

2.7.3 Summary

Two potential advantages to setting performance standards using item response theory were addressed. Both selective and adaptive standard setting allow judges to review and rate a subset of the complete item bank. In selective standard setting judges select which items they wish to rate, and in adaptive standard setting the items are selected algorithmically. Relatively little has been written about either of these procedures; however, research has shown that reasonable passing scores can be set on a subset of test items. Furthermore, studies have shown that when judges cannot omit items, passing scores are systematically lower, and less consistent. Adaptive standard setting methods have been discussed, but no empirical research has been conducted. Overall these methods appear to hold considerable promise for improving the efficiency and validity of passing scores.

2.8 Summary of the Literature Review

Overall the findings in the literature suggest that the Angoff method working on the test score scale may have significant limitations. Although judges seem to be capable of conceptualizing the minimally capable examinee, judges sometimes have difficulty placing this opinion onto the score scale. The primary limitation is that judges often seem incapable of providing internally consistent estimates of examinee performance. Furthermore, even when judges produce consistent ratings, the procedure for translating test score performance standards onto the IRT proficiency scale may distort each judge's intention.

One common practice to mitigate these issues has been the provision of examinee performance data and discussion between rounds. Although this has been shown to increase the internal consistency of judges' ratings, the practice has increasingly been called into question on the grounds that it results in a partially norm-referenced performance standard due to the use of empirical data. These results suggest that an internally consistent criterion-referenced performance standard may be unobtainable within a classical test theory framework.

The move to an IRT based standard setting method has the potential to allow judges to provide ratings for only a subset of test items. This flexibility allows passing scores to be set using either selecting or adaptive standard setting methods. In selective standard setting judges are free to omit items which they do not feel comfortable answering. In adaptive standard setting judges will respond to items which have been algorithmically selected to provide the most information in the area of the passing score. These procedures have the potential to improve both the efficiency of the standard setting process and the validity of the resultant passing score.

Despite the benefits of passing scores set using IRT the issue has been largely overlooked in the literature. The most viable procedure presented in this literature used item level theta estimates to estimate the judges' recommended passing score on the IRT proficiency scale. Although this procedure has intuitive appeal the authors use of the mean theta, is incongruous with common statistical practice. Furthermore the mean is not affected by the spread of the judge's ratings and therefore fails to reflect the inconsistency in judges' ratings.

Based on this literature review, in the next three chapters, a series of related studies designed to evaluate the feasibility and validity of passing scores set using the IRT Angoff method will be described. Chapter Three will examine the benefits of a selective standard setting procedure with specific attention to the degree to which judges provide systematically bias ratings to unfamiliar items. Chapter Four will explore the potential of an adaptive standard setting procedure to provide efficient and accurate passing scores. Finally Chapter Five will compare the stability and accuracy of passing scores set using the

IRT Angoff method as compared to the more common True Score Angoff method. Overall these chapters will provide a comprehensive view of the measurement properties of the IRT Angoff method.

CHAPTER 3

SELECTIVE STANDARD SETTING

3.1 Background and Purpose

One of the potential limitations of the True Score Angoff method is the requirement that all judges provide a rating for every item. This is because when judges provide estimates on the test score scale there is no mechanism for estimating the passing score based on a subset of item ratings. This imperative may appear to be little more than a logistical issue, but in fact it has significant implications for the validity of passing scores.

When judges are selected to serve on standard setting panels the assumption is that each judge is intimately familiar with the tested content. This familiarity allows judges to consider the knowledge and skills the examinee would need to answer the item correctly and provide a reasonable estimate of the MCE's performance. Unfortunately, at times judges lack experience with a particular content area. This may be a minor issue in K-12 achievement testing where the content domain tends to be fairly narrow, but has the potential to be a significant issue for highly technical credentialing and certification exams. For these tests the requirement that judges provide ratings for items they do not understand it has the potential to add significant errors to the estimated passing score. If these errors are presumed to be random, the effect on the final passing score may be trivial. However, logically it is easy to imagine that when judges do not feel qualified to rate the item they may perceive it as artificially difficult. If judges systematically provide lower ratings to unfamiliar items the potential exists for the passing scores to be underestimated. The purpose of this study is to determine if ratings provided for unfamiliar items are systematically lower than ratings provided for typical test items.

3.2 Research Question

This study will address one specific research question regarding the effect of unfamiliar items on judges' ratings.

1. Do individual judges produce systematically lower ratings for unfamiliar items than they do for familiar items?

This question is important because if unfamiliar items are rated systematically lower, passing scores will tend to be suppressed. The results of this study will provide evidence as to whether more valid passing scores would be obtained if judges were free to omit items.

3.3 Data

The data used in this study have been collected from an operational standard setting exercise in support of the Unites States Medical Licensing Examination (USMLE). Successful completion of the USMLE is required for all physicians with an M.D. degree seeking a license to practice medicine in the United States. The examination sequence includes three computer-delivered tests: Step 1, Step 2 CK, and Step 3. Approximately every three years standard setting exercises are conducted for each of these steps using a variation of the Angoff method on the test score scale.

The following analyses utilized standard setting data from the Step 1 and Step 2 Clinical Knowledge examinations. The Step 1 exam measures examinees mastery of the biomedical sciences and the Step 2 exam is designed to ensure that examinees possess sufficient clinical science knowledge for safe and effective care. These exams contain exclusively multiple choice items calibrated with the IRT two-parameter logistic (2PL) model. Standard setting judges for these exams are practicing physicians or non-physician PhDs working in medical education. An effort is made to recruit judges from across the

United States and to the extent possible, panels are structured to be balanced in terms of physician specialty and practice setting. The exercises are replicated three times; the replications follow identical procedures but occur on different days (typically one to two weeks apart) and include different groups of content experts.

At the start of each exercise to set a passing score content experts received orientation on the purpose of the exam. This is followed by a discussion of the knowledge, skills, and abilities possessed by the minimally proficient examinee with respect to the exam content. Judges are then instructed on the Angoff method. This discussion is followed by a group activity in which each content expert makes judgments about a practice set of 15 items. The judgments are made independently and then shared with the group. Discussion of discrepancies in ratings is focused on refining the understanding of the concept of the minimally proficient candidate. After completing this practice set of items judges review and rate 75 additional test items. Although this too is considered training, the ratings, review, and revision process fully mirrored the operational procedure. Finally, the judges provide ratings for the operational items. During this process each panel rates a unique set of items: 168 for Step 1 and 192 for Step 2. These ratings are revised in three rounds, as described earlier, with discussion and access to examinee performance data between each. The passing score resulting from the third round of operational Angoff judgments is reported to the policy group that makes the final decision about the passing score. Due to the potentially deleterious effect of examinee performance data, however, all analyses which follow will be based on the first round of expert judgments only.

One of the benefits of this data set is that while judges were pressed to provide ratings for all items, they were given the opportunity to mark items they felt uncomfortable answering. Specifically the judges were told that the examinations covered a considerable range of content and that some items may cover content with which they were “unfamiliar” or “for which they had no frame of reference for making a judgment” (M. Margolis, personal communication, March 3, 2013). If judges reviewed an item for which they did not think they could provide a sensible judgment, they were instructed to provide a judgment anyway but to mark the item by checking the box which appeared by each item. These items represented an extremely small proportion of the operational data set and therefore for the analyses presented in this chapter the 75 item training session will be used for both Steps 1 and 2.

3.4 Methodology

To determine the influence of familiarity and frame of reference on the judges' ratings separate analysis was conducted for each judge with at least one omitted item. For each of these judges, rather than relying on their Angoff ratings, all ratings were converted to the item level proficiency estimates (θij) used in the IRT Angoff method. These theta estimates are calculated using the inverse of the IRT three-parameter formula.

$$\theta_{ij} = \frac{ln \ \frac{1-c_i}{x_{ij}-c_i} \ - 1}{-a_i} + b_i$$

where a_i, b_i, and c_i are the IRT a, b, and c parameter estimates for a particular item, and x_{ij} is the Angoff rating provided by judge j on item i. Note that for the two-parameter model ci is zero for all items. The transformation onto the IRT proficiency scale not only is in keeping with the IRT Angoff method, but also eliminates the influence of empirical item difficulty on judges’ ratings.

Next for each judge all familiar items—those not marked for omission—were rank ordered based on their item level theta estimates. These ratings form the null distribution of ratings which would be expected if judges were familiar with all items. For each judge this distribution was used to identify the point on the proficiency scale which most closely approximated the 5th and 10th percentile in the lower tail. In the event that no item fell exactly on this percentile, the next lower value was selected. These points on the IRT proficiency scale served as the critical values and were used to conduct a one-tailed non-parametric test for statistical significance for each item marked as unfamiliar. When unfamiliar items fell below these critical values it indicated that the item does not belong in the null distribution. These items were therefore considered outliers. If the item level theta estimate for the unfamiliar item fell above the critical value it was considered part of the null distribution. This process was repeated for both Step 1 and Step 2.

3.5 Evaluating Results

This analysis was conducted separately for each judge, across the three panels with at least one item marked for omission. Although results are reported separately for each judge, the results in aggregate may be more telling. If judges are providing systematically bias estimates for unfamiliar items a large proportion of these items will be statistical outliers. If however, lack of familiarity with the item result in random errors only about 5% and 10% of these items respectively will be flagged as outliers.

3.6 Results

The number of unfamiliar items identified varied considerably across both exams and panels. The data for this analysis included 129 unfamiliar items in Step 1 and only twenty unfamiliar items in Step 2. Despite the discrepancy in frequency of unfamiliar items, the results across both data sets suggest that judges tend to overestimate the difficulty of unfamiliar items. Tables 2 and 3 present the number of items which fell below 0.05, between 0.05 and 0.1, between 0.10 and 0.05, and above 0.5 for each judge.

Table 2 Position of Marked Ratings Step 1

Judge	Less than 0.05	Between 0.05 and 0.10	Between 0.10and 0.50	Greater than 0.50
A	2	0	6	3
B	1	0	16	0
C	1	0	4	0
D	1	1	5	2
E	0	2	4	1
F	2	0	2	1
G	1	0	4	1
H	2	0	3	0
I	17	0	14	0
K	0	0	2	0
M	4	2	6	1
N	2	1	4	0
P	1	0	1	1
U	1	0	5	0
BB	0	0	2	0
Total	35	6	78	10

Table 3 Position of Marked Ratings Step 2

Judge	Less than 0.05	Between 0.05 and 0.10	Between 0.10and 0.50	Greater than 0.50
A	1	0	0	0
C	1	0	0	0
D	0	1	0	0

H	1	0	0	0
K	0	0	2	1
L	0	0	2	1
Q	0	0	0	1
W	0	3	1	0
X	0	0	0	3
Y	1	0	0	0
Z	0	0	1	0
Total	4	4	6	6

The results of the Step 1 analysis indicate that 35 items fell in the bottom 5% of judges' rating and 41 fell in the bottom 10%. Given a total of 129 unfamiliar items, 35 items in the bottom 5% represents approximately five times the frequency which would be expected by chance alone. Although intuitively this is a large number of outlying items, using a binomial distribution it is easy to calculate the probability of identifying this number of outlying items as a result of chance. This calculation indicates that the probability of observing this result by chance is $7.540\text{x}10^{-12}$ for 10% and $1.038\text{x}10^{-16}$ for 5%. These results strongly suggest that, at least for this examination, judges tend to systematically under predict examinees performance on unfamiliar items.

Although the number of unfamiliar items is significantly lower in the Step 2 data than in Step 1, the results indicate a similar trend. Across the 20 unfamiliar items 4 appeared in the lower 5% and 8 appeared in the lower 10%. In both cases items were placed in the lower tail of the distribution four times more frequently than would be expected purely as a result of chance. Using a binomial distribution once again, we can demonstrate that the probability of finding these results by chance is 0.0133 for the lower 5% and 0.00035 for the lower 10%. Although these results are somewhat less definitive than those found in Step 1, these findings clearly display the same systematic pattern of artificially low ratings for unfamiliar items.

These results raise the question: how would the passing scores be affected if judges were free to omit these unfamiliar items? Given that the panels' recommended passing score is based on between 675 and 825 item ratings (number of judges multiplied by 75), the impact of omitting three or four ratings will be quite minor. When a substantial number of items are marked as unfamiliar, however, the results could be extremely significant. Tables 4 and 5 present the impact of removing outlying ratings in the bottom 10% of each judge's distribution.

Table 4 Step 1 Change in Recommended Passing Score

	Original Theta	Modified Theta	Change	Original Raw	Modified Raw	Change
Panel 1	-1.71	-1.50	+0.20	40.68	43.14	+2.46
Panel 2	-1.46	-1.42	+0.04	43.70	44.16	+0.46
Panel 3	-1.84	-1.83	+0.00	39.05	39.08	+0.04

Table 5 Step 2 Change in Recommended Passing Score

	Original Theta	Modified Theta	Change	Original Raw	Modified Raw	Change
Panel 1	-1.32	-1.31	+0.01	48.88	48.99	+0.11
Panel 2	-1.51	-1.50	+0.01	47.40	47.48	+0.08
Panel 3	-0.93	-0.93	0.00	51.89	51.89	0.00

As we would expect, when very few items are omitted the impact of their removal on the recommended passing score is negligible. When a relatively large proportion of items are marked as unfamiliar, however, the impact on passing scores could be dramatic. For example Panel 1 for the Step 1 exam marked 96 items as unfamiliar. Of those 96 items 30 were in the bottom 10% of judge's ratings and 88 were in the bottom half. Table 4 shows that recalculating the passing score using only the familiar would result in the median passing score on the IRT proficiency scale increasing from -1.705 to -1.504. Using the test characteristic curve we can see that this change corresponds to a drop in the passing rate from 98% to 96.5%. Although these results would be dramatically different for other panels, these result suggest that the requirement that judges rate all items could have a practically significant and deleterious effect on the validity of the final recommended passing score.

CHAPTER 4

ADAPTIVE STANDARD SETTING

4.1 Background and Purpose

One of the advantages of item response theory over classical methods is the ability to match examinees to test items which provide the most information. This feature has allowed test developers to design adaptive exams which maximize information for each examinee. Although this adaptive approach has significant benefits for both administration time and measurement error, this procedure has never been empirically tested as an approach to setting Angoff performance standards. One of the potential benefits of the IRT Angoff method is the ability to adaptively estimate each judge's conception about the performance level of the MCE. The purpose of this study is to empirically test the accuracy and efficiency of passing scores set using an adaptive Angoff method.

4.2 Research Questions

This study will address two specific research questions designed to evaluate the accuracy and efficiency of an adaptive standard setting method.

1. Are passing scores set using an adaptive standard setting procedure comparable to the passing scores based on the complete test?
2. Are passing scores set using an adaptive standard setting procedure more accurate than passing scores based on a random sample of items?

Together these questions will be able to address the accuracy and practicality of an adaptive standard setting procedure.

4.3 Data

As with the previous study, this chapter uses standard setting results from the USMLE Step 1 and Step 2 examinations. Unlike the previous study which focused on the 75 item data sets this analysis will use the operational data set. These operational data sets contained 168 items for Step 1 and 193 items for Step 2.Theoretically the size of these data sets is more than sufficient for traditional standard setting. Therefore, rather than viewing each of these as a single standard setting exercise, the test forms will be thought of as banks of potential items. Items from these banks can then be sampled to determine what passing score would be achieved had the judges rated a particular subset of the complete item bank.

Since items will be dynamically sampled from the complete bank based on the judges' ratings, the alignment of judge opinions and item information may be informative. Figures 8 through 13 allow for the comparison of the test information function, and the density of judges' ratings. Across the six data sets included in this study, the figures show high levels of test information at the mode of the ratings density function. This alignment suggests that for all six data sets large numbers of informative items could reasonably be sampled in the area of the judges' opinions.

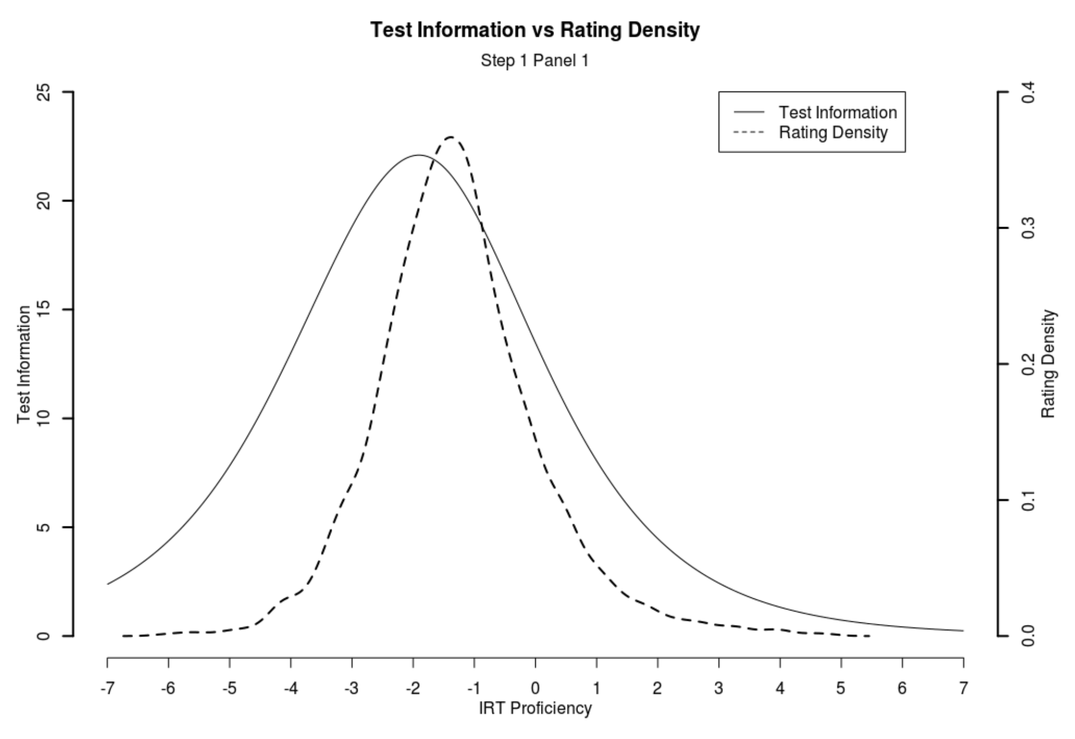

Figure 8 Step 1 Panel 1: Test Information and Ratings Density

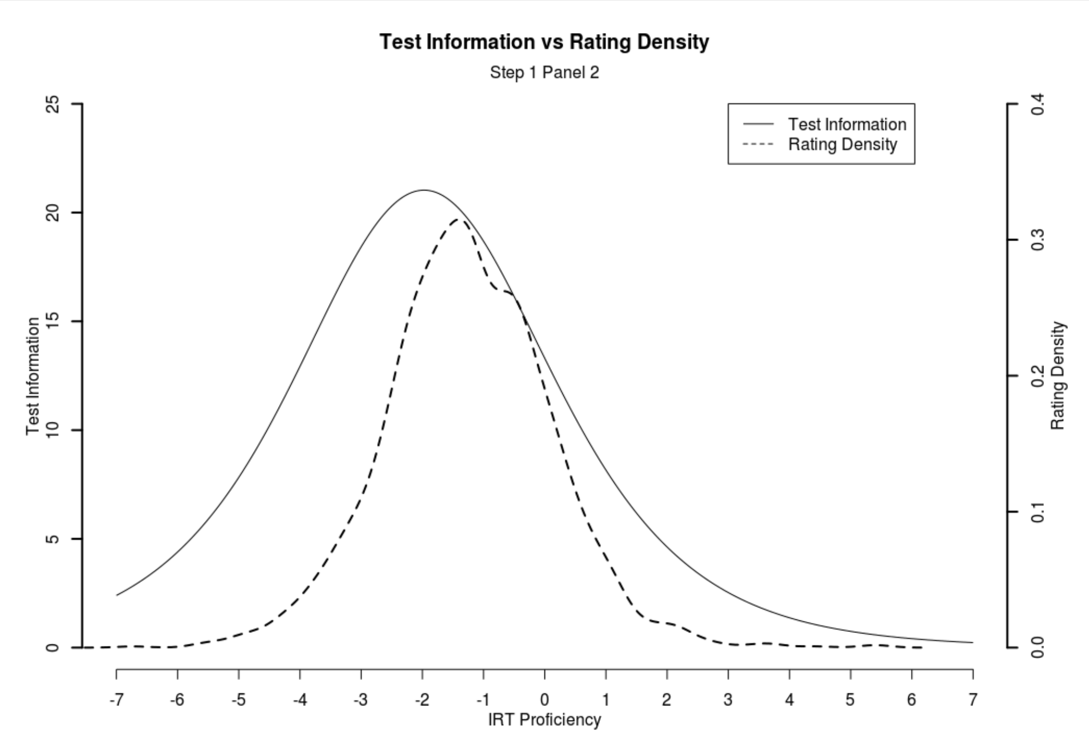

Figure 9 Step 1 Panel 2: Test Information and Ratings Density

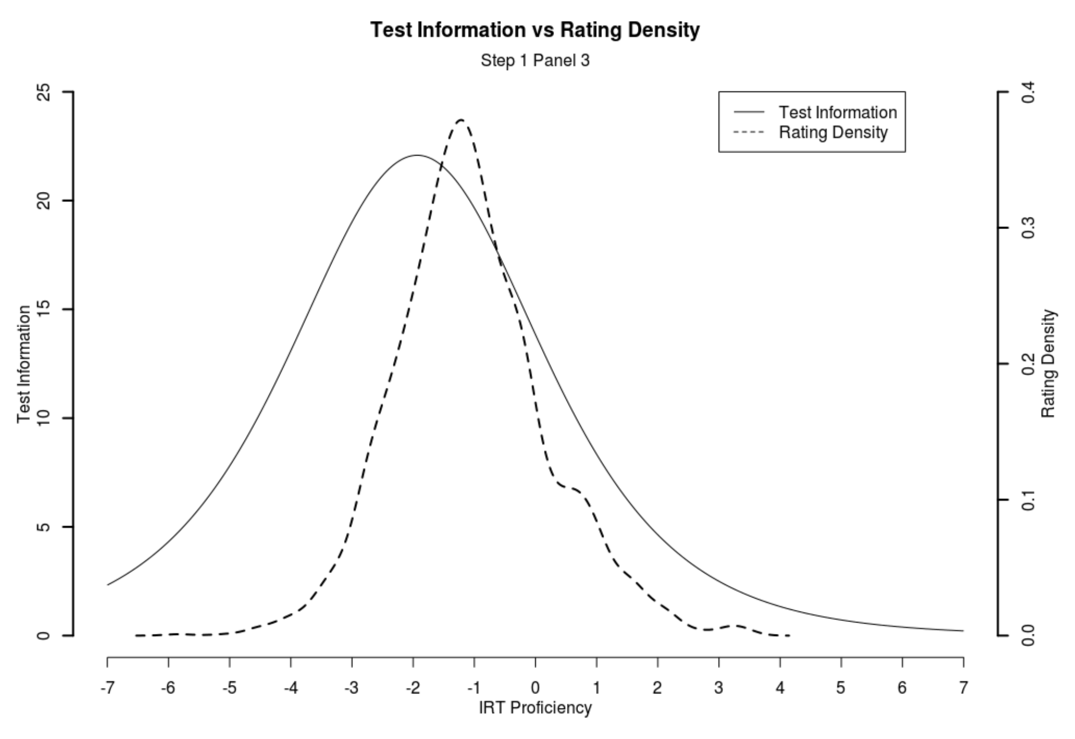

Figure 10 Step 1 Panel 3: Test Information and Ratings Density

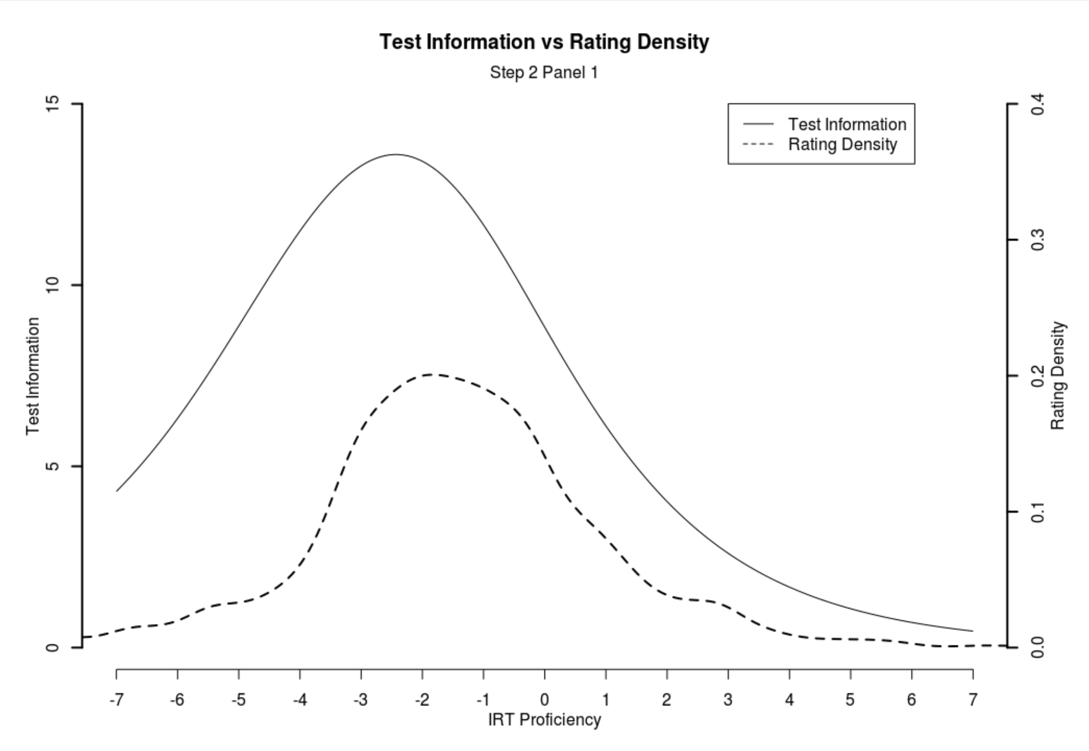

Figure 11 Step 2 Panel 1: Test Information and Ratings Density

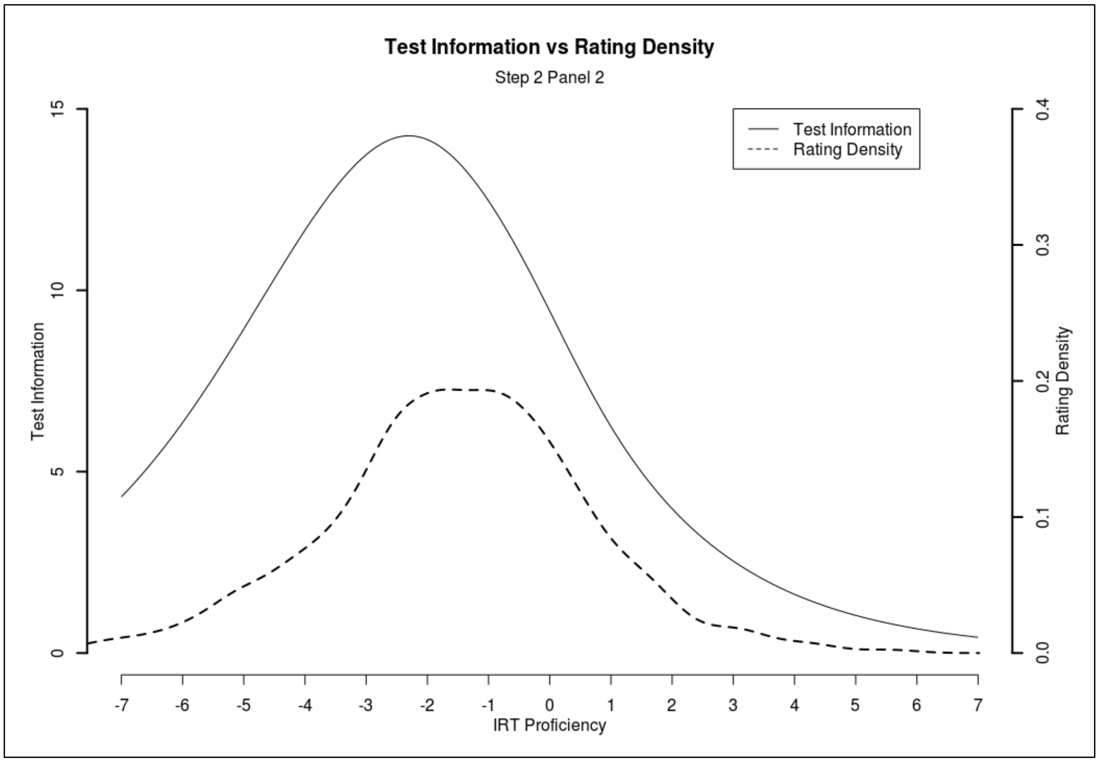

Figure 12 Step 2 Panel 2: Test Information and Ratings Density

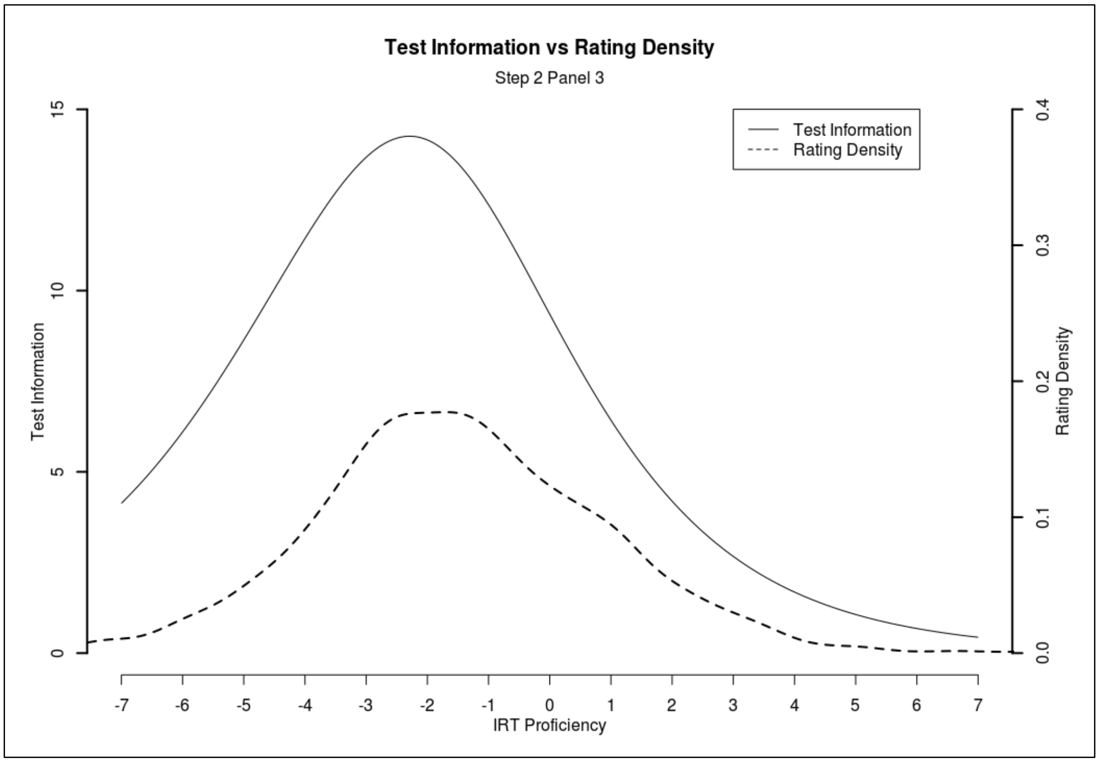

Figure 13 Step 2 Panel 3: Test Information and Ratings Density

4.4 Methodology

When developing an adaptive test three primary questions must be answered: How to Start, How to Continue, and How to Finish (Wainer, 1990). These same issues must be addressed for any adaptive standard setting method. Since this is the first empirical study of this kind, a straightforward algorithm will be applied. Additional research will be required to examine more complicated adaptive algorithms. The Starting Place, Continue, and Stopping rules will be presented in the following paragraphs.

4.4.1 Starting Place

When choosing the first item to administer in an adaptive test can be a complicated decision. Typically the first item administered would be based on the initial estimate of the examinee's ability. This estimate can be based on knowledge of past performance on the test, but often this information would not be available. In those cases the first item is typically chosen based on the mean ability in the population of examinees. In the case of content experts setting passing scores, the average opinion in the population of judges is unknown. In fact, it is this very value that the standard setting method is attempting to estimate. Therefore rather than starting with an item which mirrors our estimate of the judge's initial opinion, each judge will begin by rating the median difficulty item within the bank. Although this item is not necessarily ideal, it is presumably more appropriate than an item selected at random.

4.4.2 Continuing

After the first item is selected rules must be developed for selecting the next item to be administered. In a typical adaptive test, items are selected based on the current estimate of the examinees ability. In adaptive standard setting the same logic applies, but the estimation of the judge's opinion is based on the IRT Angoff method. In the IRT Angoff method each rating will be converted into an item level theta estimate using the formula presented in Chapter Three. When the judge rates the first item, the theta associated with that item will be the initial estimate of the judge’s internalized ability. The next item selected will be the one which has the most information at the current estimate of the judge's opinion of the MCE's ability. The current estimate of the judge’s opinion is the median of the item level ratings for the items administered to that point. Ability estimates will be refined and items will be continually selected until the stopping rule has been reached.

4.4.3 Stopping

In typical adaptive testing environments examinees continue to see items until the standard error of their ability estimate reaches an acceptable level, or some predefined number of items has been administered. In adaptive standard setting either of these criteria could reasonably be employed, however, since standard setting activities are completed as a group there is relatively little benefit to having individual judges seeing substantially fewer items than their peers. Since the judge cannot move on without the rest of the panel it seems reasonable to have all judges respond to the same number of items. Therefore in this study seven different stopping conditions were examined. The panel recommended passing scores to be calculated based on 15-75 items in increments of 10. Although these stopping points are arbitrary, 75 items test has been chosen to reflect the length of the full standard setting training exercise. This exercise is often treated as a complete test and therefore seemed appropriate to consider a full length test for the purposes of this simulation. The shorter tests from 15-65 indicate how efficiently passing scores can be set using the adaptive algorithm.

4.5 Evaluating Results

4.5.1 Comparison to Truth

To evaluate the accuracy of adaptive performance standards across the seven conditions it was necessary to compare the adaptive cut score to the cut score which would have been achieved had the passing score been set using traditional methods. This could mean comparing the adaptive cut score to the single passing score based on the complete bank of items, but since the passing scores exists on a continuous scale it is effectively impossible to observe the identical passing score based on only a subset of items. Instead a distribution of acceptable passing scores was calculated based on a 150 item test. The decision to use 150 items seemed appropriate since many unique samples can be drawn from the bank and the 150 items is double the length of the full 75 item test. Therefore all passing scores in the distribution of acceptable passing scores will contain significantly less error than the 75 item exam. For each judge a sample of 150 items was selected at random and the panel level passing score was calculated using the IRT Angoff method described above. This process was repeated 1,000 times, each time drawing a new sample of 150 items for each judge. The distribution of cut scores makes it possible to calculate the probability of finding the adaptive performance standard on the 150 item test. If the adaptive performance standard would occur with a high degree of regularity, it may indicate that the adaptive performance standard is a reasonable replacement for the passing score set on all items. If the adaptive performance standard is an outlier relative to the distribution of standards set on all items, it would not be a suitable replacement.

4.5.2 Comparison to Random Sample

In addition to comparing the adaptively set passing scores to a distribution of acceptable replacements, this study evaluated the probability of finding a more accurate passing score from a random sample of items. Although theoretically, adaptive standard setting should result in greater accuracy with fewer items this may not be the case in practice. This analysis compared the observed adaptive passing score for each condition to a distribution of passing scores set using an equal number of randomly selected items. These random samples were drawn 1,000 times for each condition and the mean absolute difference (MAD) between these observed passing scores and the overall passing score based on all items was calculated. The magnitude of the MAD for the adaptive passing score was compared to this null distribution. If observed adaptively set passing scores are below the 5th percentile of the null distribution it indicates that they are more accurate than the results which would be expected by chance. All greater values indicate that more accurate passing scores would frequently be found simply by drawing a random set of items for each judge.

4.6 Results

4.6.1 Comparison to Truth

Adaptively set passing scores were compared to a null distribution based on 1,000 random samples of 150 items for each panel in both data sets. Scores which fell between the 5th and 95th percentile of this null distribution for each panel were defined as acceptable passing scores. Any observed passing score falling outside this range was considered unacceptable for the purposes of this study. Tables 6 and 7 present the range of acceptable passing scores across each of the three panels across both data sets.

Table 6 Range of Acceptable Passing Scores Step 1

	5%	All Items	95%	Range
Panel 1	-1.380	-1.361	-1.340	0.040
Panel 2	-1.303	-1.264	-1.261	0.042
Panel 3	-1.169	-1.162	-1.141	0.028

Table 7 Range of Acceptable Passing Scores Step 2

	5%	All Items	95%	Range
Panel 1	-1.537	-1.488	-1.445	0.092
Panel 2	-1.587	-1.531	-1.453	0.134
Panel 3	-1.623	-1.579	-1.521	0.103

The results indicate that the simulation has produced a narrow null distribution centered around the observed passing score based on all operational items for each panels and data sets. Using this information the observed adaptively set passing scores can be compared to these distribution of acceptable standards. Tables 8 and 9 present adaptive passing scores for Step 1 and Step 2 respectively across all panels and stopping rule conditions.

Table 8 Adaptive Passing Scores Step 1

	15 Items	25 Items	35 Items	45 Items	55 Items	65 Items	75 Items
Panel 1	-1.635	-1.597	-1.591	-1.518	-1.423	-1.421	-1.428
Panel 2	-1.366	-1.251	-1.234	-1.238	-1.253	-1.263*	-1.263*
Panel 3	-1.161*	-1.177	-1.135	-1.124	-1.135	-1.154*	-1.160*

* Indicate a passing score falling within the acceptable range

Table 9 Adaptive Passing Scores Step 2

	15 Items	25 Items	35 Items	45 Items	55 Items	65 Items	75 Items
Panel 1	-1.696	-1.696	-1.702	-1.702	-1.792	-1.768	-1.696
Panel 2	-1.726	-1.773	-1.850	-1.814	-1.740	-1.700	-1.608
Panel 3	-1.861	-1.423	-1.423	-1.549*	-1.598*	-1.616*	-1.598*

* Indicate a passing score falling within the acceptable range

The results indicate the three of the six panels produced passing scores in the acceptable range with 65 or fewer items. These results, while potentially encouraging, are far from definitive as the other three panels produced passing scores outside the acceptable range. It is interesting to note that the observed passing scores were incredibly consistent across the seven conditions. This would seem to indicate that selecting items based on item information tends to produce reliable standards even when only a modest number of items are selected. Although these results make it difficult to draw any strong conclusions it may be reasonable to conclude that under certain circumstances adaptive standard setting could produce accurate passing scores with a fraction of the total items administered.

4.6.2 Comparison to Random Sample

Results of the previous analysis indicate that adaptive standard setting may at times provide an acceptable alternative to traditional methods. These findings do not, however, indicate the probability of observing these results by chance with a randomly selected sample of items. Tables 10 and 11 indicate the probability of observing a more accurate passing score when items are selected at random. Values lower than 0.05 indicate that the adaptive standard is significantly better than chance, while all greater values suggest that the adaptively set passing score would be observed by chance with a fairly high degree of regularity.

Table 10 Probability of Observing More Accurate Passing Score Step 1

	15 Items	25 Items	35 Items	45 Items	55 Items	65 Items	75 Items
Panel 1	0.483	0.593	0.709	0.756	0.844	0.894	0.930
Panel 2	0.030*	0.074	0.088	0.083	0.125	0.132	0.159
Panel 3	0.019*	0.028*	0.035*	0.041*	0.047*	0.065	0.067

* Indicate a statistically significant finding

Table 11 Probability of Observing More Accurate Passing Score Step 2

	15 Items	25 Items	35 Items	45 Items	55 Items	65 Items	75 Items
Panel 1	0.674	0.806	0.887	0.936	0.999	1.000	0.991
Panel 2	0.610	0.864	0.975	0.968	0.938	0.857	0.616
Panel 3	0.844	0.648	0.754	0.240	0.180	0.333	0.239

* Indicate a statistically significant finding

When interpreting these results it is important to remember that results across the seven length conditions are not independent. It is therefore inappropriate to aggregate results across conditions. It is, however, possible to compare results within each condition. These results suggest that, with the exception of the Step 1 Panel 3 results, the adaptive passing scores do not offer significant improvement over selecting items at random. At times, as is the case with Step 2 Panel 1, the adaptive results are significantly worse than what could reasonably be expected by chance. Together these findings would seem to suggest that selecting items based on information influences passing scores, but does not yield passing scores which are comparable to those based on all items. Rather than simply selecting an ideal set of test items, the adaptive procedure appears to select items which have the potential to bias Angoff ratings. This results in reliable standards across conditions, but does not necessarily lead to more accurate standards and accuracy is a more important criterion than stability or reliability.

CHAPTER 5

STABILITY OF PERFORMANCE STANDARDS

5.1 Background and Purpose

This study will examine the degree to which item samples affect the stability and accuracy of estimated passing scores using the IRT Angoff method. The focus on both stability and accuracy in this study will represent a modest departure from the majority of the literature on this topic. When working with operational data it is typically not possible to compare passing scores to some known truth. Instead the stability of the passing scores has served as an important source of validity evidence. Although this approach is quite reasonable, it is only effective in identifying random errors in the estimate of the passing score. Any systematic error introduced by the standard setting method, would result in a shift in the mean passing scores without influencing the stability.

This study will examine the effect of both systematic and random error on the estimation of the passing score across different pools of items. Specifically random errors are produced when judges struggle to provide consistent estimates of examinee performance across items. This inconsistency will result in random variability in the recommended passing score across sets of items. The systematic error may be introduced by the mapping of test score performance standards onto the IRT proficiency scale. This issue was discussed in greater detail in section 2.5, but in general, the TCC mapping approach to placing the test score performance standard on the IRT proficiency scale will produce passing scores which are influenced by the selection of items. This effect may not influence the stability of the estimated passing score but has the potential to systematically bias the resulting passing score. The purpose of this study is to determine if passing scores developed using the IRT Angoff method can improve the stability and accuracy of passing scores developed using the True Score Angoff method.

5.2 Research Questions

This study will address three specific research questions regarding both the stability and accuracy of Angoff passing scores.

1. Does the IRT Angoff method produce more stable passing scores than the True Score Angoff method?
2. Are True Score Angoff passing scores systematically related to mean item difficulty?
3. Does the IRT Angoff method produce estimated passing scores closer to the known true passing score than the True Score Angoff method?

Each of these research questions is focused specifically on the stability and accuracy of Angoff performance standards. The results of this study will help to determine if an IRT-based approach to the Angoff method has the potential to yield more valid passing scores by improving the stability and accuracy of estimation.

5.3 Data

Because this study is concerned with comparing observed performance standards to a known true value, both true performance standards and individual ratings were simulated. Despite the use of simulated data, the simulation relied on realistic item parameter estimates and ratings based on actual judge behavior. The simulated judges and ratings are designed to reflect the true distribution of judges and ratings in the USMLE Step 2 75 item data set. Basing the simulated data on this observed data set, helps to ensure that results found in the simulation could reasonably be observed as part of an operational setting activity.

5.4 Methodology

To examine the stability and accuracy of Angoff performance standards a simulation study will be conducted. Although simulation studies are fairly uncommon in standard setting research, in this case a simulation study was chosen so that observed recommended passing scores could be compared to a known true passing score. If this study had relied exclusively on operational data it would have eliminated many of the assumptions which accompany simulation research, however it would have eliminated our ability to compare observed passing scores to truth. The simulation consisted of four parts: sampling of items, judges, ratings, and calculation of both IRT and True Score Angoff passing scores. The full simulation was conducted seven times with different mean item difficulties and each condition consisted of 1,000 replications. The details of the simulation follow.

5.4.1 Sampling of Items

For this study item parameters were based on the three-parameter model. For each replication 75 three-parameter items were randomly selected. The a-,b-, and c-parameters were selected based on the recommended prior distributions used in Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). The parameters were randomly drawn from the following distributions.

a-parameter: The discrimination parameter is the natural logarithm of the values sampled from a normal distribution with a mean of 0.0 and a standard deviation of 0.5. This will result in a-parameters with a mean of 1.0 and a lower bound at 0.0.

b-parameter: The difficulty parameter was drawn from a normal distribution with a mean of 0.0 and a standard deviation of 1.0.

c-parameter: The pseudo-guessing parameter was drawn from a beta distribution with parameters at 5 and 17. This resulted in a positively skewed distribution with a mode at 0.2 bounded between 0.0 and 1.0.

This process generated reasonable three-parameter items on the familiar (0,1) scale.

Next, to examine the potential bias introduced by item difficulty on the Angoff passing scores, the mean item difficulty must be manipulated. In this study a wide range of mean item difficulties were used to demonstrate the effect of difficulty on the position of observed passing scores. To understand this effect seven mean item difficulties were considered. These difficulties were at seven integer values between -3.0 and 3.0. This design will not only illustrate if mean item difficulty affects passing scores, but will also help to determine if the magnitude of this effect is systematically related to the distance between mean item difficulty and the true passing score.

Before these item parameters were used as part of the simulation study, the item parameters and judges internalized thetas must be equated onto the same scale. In principle two scaling approaches could be employed: either simulated item parameters could be brought onto the USMLE scale, or judges ratings could be brought onto the 0,1 scale. Although there are advantages to both approaches, in this study judges' ratings will be brought onto the 0,1 scale so that the interpretation of the results will be more intuitive. This transformation was done using mean-sigma equating. Mean-Sigma equating is a common form of linear equating used in Item Response Theory. Mean-Sigma equating is based on the principle that the standard deviate score on both scales should be equal.

$$\frac{b_{random} - b_{random}}{SD(b_{random})} = \frac{b_{USMLE} - b_{USMLE}}{SD(b_{USMLE})}$$

To convert the USMLE theta values onto the 0,1 scale we can rearrange this formula into slope intercept form.

$$b_{USMLE} = \frac{SD(b_{USMLE})}{SD(b_{0,1})} b_{0,1} - \ b_{random} - \frac{SD(b_{USMLE})}{SD(b_{0,1})} b_{0,1}$$

Therefore the observed theta values on the USMLE can be converted to the 0,1 scale as follows.

$$\theta_{USMLE} *= m\theta_{0,1} + n$$

where m is the slope and n is the intercept from the above equation. This procedure equates all USMLE theta scores onto the common 0,1 scale.

5.4.2 Sampling of Judges

After item parameters have been selected a simulated sample of content experts must be selected. For this study we stipulated that the true passing score is the mean passing score which would be obtained by sampling the opinions of all qualified content experts. Although it is expected that individual judges will deviate somewhat from this mean, the magnitude of the deviation can be approximated using a normal distribution centered around the group mean. To establish reasonable values for each judge, the distribution of opinions was based on the equated values from the USMLE Step 2 examination. Across the three panels of observed ratings the mean proficiency level was -1.314 with a standard deviation of 0.986. To facilitate interpretability the distribution mean was shifted to 0.000. Therefore each judge's belief regarding the ability of the MCE was drawn from a normal distribution with a mean of 0.00 and a standard deviation of 0.986. For each replication a new panel of 10 simulated judges were drawn from this distribution, but the true performance standard of 0.00 will be consistent across all replications.

5.4.3 Sampling of Ratings

Although this simulation assumes that each judge has a single opinion of what ability level is required for an examinee to be considered minimally proficient, it is unrealistic to assume that judges will be perfectly consistent across all ratings. To reflect this inconsistency each judge's individual item ratings was drawn from a normal distribution centered around his or her true mean. Although this instability was simulated on the IRT proficiency scale it represents errors which result both from changing views of the MCE's ability and errors in the estimation of examinee performance. The standard deviation of each judge’s error distribution will once again mimic the equated ratings from the USMLE Step 2 data. The standard deviation of a judge’s ratings was drawn from a uniform distribution bounded by 1.459 and 3.523. The uniform distribution was selected to reflect the spread of standard deviations seen in the operational data and to ensure that the standard deviations do not become negative. During each replication each judge will receive a set of 75 item level theta estimates centered around the judge's known true mean.

5.4.5 Calculation of Passing Scores

During each replication three passing scores were calculated: the first two using the IRT Angoff method and a third using the True Score Angoff method. The calculation of each of these passing scores is fairly straightforward, but it is complicated somewhat because rather than getting judge's ratings, and calculating the corresponding item level theta, the process is reversed. In this simulation we begin with a theta for each item and use item response theory to calculate the corresponding conditional p-values. These initial thetas and the corresponding calculated p-values are then used to calculate each passing score.

5.4.6 IRT Angoff

The calculation of the IRT-Angoff passing scores was made significantly easier in this simulation because each judge begins with a set of 75 item level theta estimates. Two passing scores were calculated using this data. The first is simply the median and the second is simply the mean of the theta estimates across all judges and items. These two IRT-based Angoff passing scores will be saved after each replication to determine the stability of passing scores set using this method.

5.4.7 True Score Angoff

The calculation of the True Score Angoff method begins with each judge's set of 75 item level theta estimates and the 75 simulated test items. To calculate the judges expected rating on the item the IRT three-parameter formula is used.

$$P = c_i + (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}$$

Given an ability and item parameters this formula produces a probability estimate. This probability is calculated for all judges across all item ratings. This results in 75 conditional p-values for each of the 10 simulated judges. These ratings are summed across judges to produces a passing score on the test score scale for each judge and then averaged across the panel to produce a recommended passing score for the panel.

Next the test characteristic curve (TCC) is calculated for the random sample of 75 items. The TCC is the sum of the individual item characteristic curves. Mapping through the TCC produces a single passing score on the proficiency scale for each possible passing score on the test score scale.

$$T|\theta = \sum_{i=1}^{75} P_{\mathrm{i}}(\theta)$$

Where T is the expected test score, at a given ability level, and Pi is the probability of a correct response on item i for a person with that ability level. The point on the proficiency scale associated with the recommended passing score on the test score scale was saved for each replication as the Angoff passing score.

5.5 Evaluating Results

This simulation resulted in three sets of 1,000 passing scores (two others using an IRT-based Angoff and one other based on True Score Angoff) for each of seven item difficulty conditions. These results must be evaluated to answer our three research questions. The evaluation procedures will be described below for each of these questions.

Does the IRT Angoff method produce more stable passing scores than the True Score Angoff method?

The stability of passing scores was evaluated based on the spread of the passing scores across the 1,000 replications for each set of conditions. The spread of the data was determined using the standard deviation of passing scores across the 1,000 replications. The standard deviation is appropriate because the simulation assumes that all errors are normally distributed. Therefore the resulting distribution of passing scores is expected to be symmetric. Furthermore, using the standard deviation allows for estimates of the probability of observing particular passing score. The standard deviation across the three sets of passing scores was calculated for each of the seven conditions. The method with the lowest standard deviation can be said to be the most stable.

Are True Score Angoff passing scores systematically related to mean item difficulty?

To determine the effect of item difficulty on Angoff performance standards the mean passing score for each method was considered across the seven difficulty conditions. If item difficulty does not affect the passing score the mean passing score should be consistent across the seven conditions. If mean passing scores are systematically linked to item difficulty the mean passing score should be related to the direction and magnitude of the shift in mean item difficulty.

Does the IRT Angoff method produce estimated passing scores closer to the true passing score than the True Score Angoff method?

This question is ultimately the most important because it combines the consistency and accuracy of each standard setting method. To determine the average distance from the true passing score the mean absolute difference (MAD) between the observed passing score and the known true performance standard of zero was calculated. The mean absolute difference is the average discrepancy between true and simulated passing score when all residuals are made positive.

$$MAD = \frac{{}_{r=1}^{1,000} \mid S_r - 0 \mid}{1,000}$$

In the formula above r is the current replication and Sr is the observed passing score for that replication. This statistic was selected because it is not affected by the direction of the error and results are reported on the IRT proficiency scale. The MAD was calculated one time for each difficulty condition. The method with the smallest MAD under each condition was considered the most accurate.

5.6 Results

5.6.1 Stability of Observed Passing Scores

The simulation generated three sets of 1,000 passing scores for each of seven item difficulty conditions. Since all replications and conditions were simulated with the same known true passing score, the variation of the simulated passing scores cannot be attributed to some changes in judges' opinions. Instead all variability observed in the simulated passing scores can be attributed to sampling error in the selection of judges or lack of internal consistency in the judges' simulated rating. Although the magnitude of these two sources of error will profoundly affect the stability of the recommended standard in each replication, three cut scores were based on exactly the same judges and ratings. It is therefore reasonable to attribute differences in the stability of ratings, solely to the selection of standard setting method.

The variability of the simulated standards was captured using the standard deviation across the 1,000 replications for each standard setting method and condition. Results from this analysis are presented in Table 12.

Table 12 Standard Deviation for Passing Scores at Each Difficulty Condition

	-3	-2	-1	0	1	2	3	Range	Average
Median IRT	0.289	0.292	0.290	0.295	0.294	0.287	0.301	0.013	0.293
Mean IRT	0.275	0.275	0.278	0.282	0.283	0.271	0.288	0.017	0.279
True Score	0.250	0.219	0.204	0.202	0.211	0.216	0.260	0.058	0.223

The results indicated that for the IRT-based methods the stability of recommended standards is fairly consistent across the seven difficulty conditions. In fact, within either of these standard setting methods the largest difference between any two conditions is 0.017. For traditional Angoff method, however, the stability of standards seemed to be systematically influenced by the mean item difficulty. Across the seven conditions the least variability was seen in mean difficulty of the zero condition with greater variability observed as mean item difficult deviated from zero. Although the effect was fairly modest the systematic nature of the errors lead to a range across the conditions of 0.058.

In addition to this pattern observed across conditions, clear patterns emerge in the stability across methods. This effect is most clearly observed by comparing average standard deviations across the three standard setting methods. Here the IRT-based approaches result in a mean standard deviation of 0.293 and 0.279 for the approach using the median and mean theta respectively. The traditional Angoff method on the other hand was considerably more stable with a mean standard deviation across the seven conditions of 0.223. Not only is the average lower across the seven conditions but, the widest spread observed in the traditional Angoff method (0.260) is lower than the narrowest spread observed in either of the IRT-based methods (0.271).

Although overall these differences in the spread of recommended passing score may appear very minor, the difference could have a profound effect on observing aberrant passing scores. For example when the mean item difficulty is zero the likelihood of observing a passing score more than 0.5 above or below the mean is only 1.3% for the traditional Angoff method. For the IRT-based methods the likelihood is 9.0% and 7.6% for the median and mean methods respectively. Overall these results would seem to indicate that the traditional Angoff method will produce the most consistent results across replications.

5.6.2 Impact of Item Difficulty on Observed Passing Scores

To evaluate the impact of mean item difficulty on observed passing scores, mean passing scores were compared across method and condition. In this simulation the known true passing score was set to zero on the IRT proficiency scale for all conditions. Systematic deviations from this known true standard across a method may indicate that observed standards are not invariant to item selection. Table 13 presents the mean observed standard across the 1,000 replications for each difficulty condition and standard setting method.

Table 13 Mean Passing Scores at Different Difficulties Conditions

	-3	-2	-1	0	1	2	3
Median IRT	-0.014	-0.004	-0.006	-0.007	0.013	0.020	0.007
Mean IRT	-0.019	0.000	-0.002	-0.005	0.016	0.017	0.003
True Score	-0.872	-0.603	-0.315	-0.002	0.323	0.619	0.860

The results in Table 13 indicate that the simulated passing score across the IRT-based standard setting methods appear to be unaffected by mean item difficulty. Across these methods, the passing scores are consistently close to 0.0, with the range in average passing score of -0.019 to 0.020. Furthermore even these modest deviations from the known true passing score appear to be random across difficulty conditions.

The traditional Angoff method on the other hand exhibits a large and systematic bias in the simulated passing scores. When the mean item difficulty is equal to the known true standard no bias is observed, but as item difficulty and the true passing score diverge observed passing scores shift in the direction of the item difficulty. These errors appear to be linearly related to the mean item difficulty, with a shift in observed passing score equal to about 30% of the shift in mean difficulty. For example a shift in item difficulty of 1.0 results in a change in passing score of approximately 0.3. These results strongly suggest that passing scores set with the traditional Angoff method are not invariant across test items.

5.6.3 The Accuracy of Observed Passing Scores

The final stage of this study brings together the stability and position of the simulated passing scores to evaluate overall accuracy. In this study accuracy was represented by the mean absolute difference between the true and observed passing scores. Table 14 presents the mean absolute differences across difficulty conditions and standard setting method.

Table 14 Mean Absolute Difference between True and Observed Passing Scores

	-3	-2	-1	0	1	2	3	Average
Median IRT	0.234	0.235	0.233	0.238	0.234	0.232	0.243	0.236
Mean IRT	0.222	0.221	0.219	0.225	0.226	0.218	0.231	0.223
True Score	0.872	0.604	0.327	0.160	0.335	0.619	0.860	0.540

Across the two IRT-based methods the mean absolute differences between true and observed passing scores are fairly consistent. These errors do not appear to be systematically linked to item difficulty and can largely be attributed to lack of stability in the observed standards. The results for the traditional Angoff method, on the other hand show considerable variability in the mean absolute difference. These errors are a function of both the variability in the observed passing scores and the systematic bias introduced by item difficulty. Figure 14 demonstrates the distribution of observed True Score Angoff passing scores for the seven difficulty conditions.

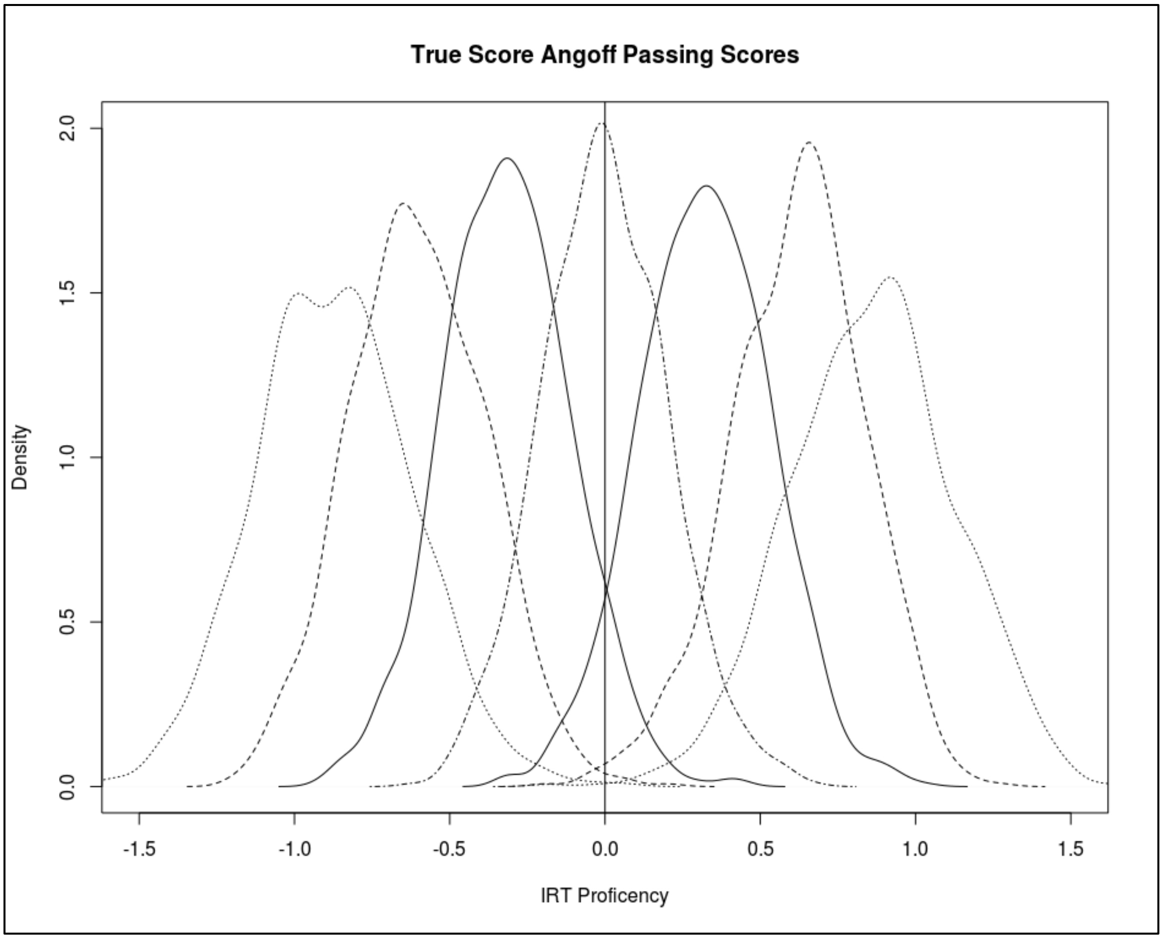

Figure 14 Distribution of True Score Angoff Passing Scores

From this figure it is clear that the bias introduced by the different item difficulties can dramatically undermine the accuracy of the standard. When item difficulty is separated from the judges true opinion by +/-2 the likelihood of observing a standard within 0.5 of truth is approximately 35%. This likelihood falls to approximately 7% when item difficulties deviate by +/-3. Overall these results suggest that the traditional Angoff method has the potential to produce significant systematic errors, when judges' opinions and item difficulty are not well aligned.

CHAPTER 6

DISCUSSION

6.1 Introduction

The Angoff method is the most popular standard setting procedure used in certification and licensure testing. Like all standard setting procedures the Angoff method is a systematic procedure for placing expert opinion on the score scale. In the Angoff method content experts review all test items and provide an estimate of the proportion of minimally competent examines who would answer each item correctly. These ratings are then summed across items and averaged across judges to arrive at the panels recommended cut score on the test score scale. Although the Angoff method is straightforward and theoretically appealing previous research has identified several limitations to the implementation of the method in practice.

One of the primary threats to the validity of the Angoff method is the inability of content experts to produce internally consistent ratings. Although judges' professional experience may allow them to internalize the ability of the MCE, research has suggested that they often struggle to estimate examinee performance on specific items. When judges fail to produce internally consistent Angoff judgments the individual ratings do not correspond to a single point on the ability scale. At times these ratings may suggest that examinees of dramatically different ability levels are all "minimally proficient". When these judge-level errors are aggregated into a single point on the ability scale the overall recommended passing score may fail to reflect the true opinions of the content experts.

In addition to a judge's limited ability to estimate examinee performance, the common Angoff method is further limited by the potentially item dependent nature of the resulting passing score. Although the judges' belief regarding the ability of the MCE is theoretically independent from item difficulty, this is not strictly true in practice. Instead the specific scale transformation typically used in the common Angoff method has the potential to results in different passing scores solely as a result of item selection. Since peculiarities of the scale transformation have the potential to influence the final recommended passing score, this limitation means that passing scores developed using the traditional Angoff method will not properly reflect the opinions of the content experts.

The final limitation of the common Angoff method as presented in this study is the requirement that all judges provide ratings for all items. This requirement is not only inefficient but also has the potential to bias the recommended passing score. This bias may occur when judges are required to rate items they do not fully understand. If the rating provided for these items are systematically different than other item ratings, the resulting passing score will be artificially influenced by the judges unfamiliarity with the tested content. Although the magnitude of these errors will depend on the expertise of the panel relative to the tested content area, at times these errors have the potential to have a practically significant effect on the recommended passing score.

The goal of the Angoff method is to infer the opinions of content experts and to reflect that opinion as a point along the score scale. Unfortunately the common Angoff method has several significant limitations which interfere with its ability to properly estimate the true opinions of the content experts. Several of these limitations are due to the fact that the Angoff method is grounded in classical test theory. Therefore interpreting Angoff ratings within an item response theory framework could theoretically mitigate these limitations. This study has attempted to extend the previous research to comprehensively evaluate the properties of Angoff standards developed wholly within an item response theory framework. Therefore the purpose of this study has been to examine the benefits of using IRT to interpret Angoff ratings. With this in mind, three studies have been presented which empirically evaluate the benefits of conceptualizing Angoff method within an IRT framework.

The balance of this chapter will discuss the implications of these studies and present recommendations for future research. The chapter has been divided into the following five sections:

1. Selective Standard Setting. This section will discuss the degree to which requiring all judges rate all items will influence overall recommended passing scores.

2. Adaptive Standard Setting. This section will outline the potential efficiency offered by algorithmically selecting items in the Angoff standard setting method.

3. Stability of Performance Standards. This section will evaluate the degree to which an IRT-based approach to Angoff results in a more stable and accurate estimated passing score than the common Angoff method.

4. Overall Evaluation. This section will summarize results from the three studies to evaluate the benefits of conceptualizing the Angoff method within an IRT framework.

5. Limitations and Future Research. This section will highlight the limitations of this study and will suggest areas where future research is needed.

6.2 Selective Standard Setting

When selecting content experts to participate in an Angoff standard setting panel it is critical that all judges are intimately familiar with the tested content. This familiarity allows judges to consider the knowledge and skills the minimally competent examinee would need to answer the item correctly and provide an informed estimate of his or her performance. Even when eminently qualified judges are selected, small gaps in their content knowledge may be unavoidable. This has the potential to become a particularly significant problem in certification and licensure testing where the domain of content is often both broad and highly technical. Unfortunately when judges are required to provide ratings for these unfamiliar items, as they are in the common Angoff procedure, this content knowledge deficit may systematically suppress the passing score. This limitation is easy to understand. When content experts are asked to estimate the performance of the MCE they must evaluate the difficulty of the item. If the judges are unfamiliar with the items tested content, the judge may overestimate the difficulty of the item and therefore underestimate examinee performance. When this pattern occurs across many judges and items the effect would be to artificially suppress the passing score. Although theoretically this limitation could seriously influence the passing score, this effect has not been demonstrated in practice. Historically, panelists have not been given the option of omitting test items and the topic has not received much attention from researchers. This study evaluated actual standard setting data to determine if judges provide systematically different rating for unfamiliar items, and if so, to what degree will this issue influence the overall recommended passing scores.

6.2.1 Summary of Results

The purpose of this study was to determine if judges provide systematically different ratings for familiar and unfamiliar items. The results of this study indicate that across both Step 1 and Step 2 of the exam the judges' ratings for unfamiliar items were systematically lower than for familiar items. Specifically the results indicate a disproportionately large number of unfamiliar items were placed in the bottom 5% and 10% of the distribution. This frequency of unfamiliar items was dramatically greater than would be expected simply as a result of chance indicating a statistically significant difference in the ratings judges provide for familiar and unfamiliar items. These results directly support the hypothesis that judges overestimate the difficulty of unfamiliar items. The results in Tables 2 and 3 clearly show that the vast majority of unfamiliar items appear in the lower half of the distribution. Furthermore, no items appeared in the upper 5% tail and only 1 item appeared in the upper 10% tail. These results suggest a systematic pattern to the judges' ratings rather than merely an increase in random error. Although these results may not generalize to all judges or standard setting environments, these results clearly show that in this specific context judges tend to provide systematically lower Angoff ratings to items they identify as unfamiliar.

Although these findings are compelling, these results do not ensure that the impact on passing scores will be of any practical consequence. To understand the practical significance of these results the passing scores were calculated two ways. The first method simply used the IRT Angoff method with all items used to calculate the passing score. The second method was designed to mimic a scenario in which judges were free to skip unfamiliar items. This approach calculated the passing score using the IRT Angoff method but omitted all unfamiliar items. The results of this calculation indicated that for five of the six panels the passing score increased when unfamiliar items were omitted and one panel showed no change in the passing score. However, only one panel saw a large enough change in the overall recommended passing score to shift the raw passing score by even one-half of a point. These results suggest that, unsurprisingly, the influence of unfamiliar items on passing score is extremely dependent on the proportion of rated items which are unfamiliar to the content experts. When content experts are familiar with virtually all tested content, the systematic bias has an extremely small impact on the actual passing score. When a large portion of items are unfamiliar the change in passing score could have a significant influence on the final recommended passing score.

6.2.2 Concussions

These findings call attention to the large discrepancy in the actual number of items marked as unfamiliar by each panel. Across the six panels of approximately equal size, the number of marked items ranged from six to ninety-six. The analysis presented in this study did not specifically examine the sources of this variability; however this peculiarity cannot be ignored completely given its potential to influence passing scores. The following paragraphs will consider possible sources of this variability and discuss the implications for the interpretation of passing scores.

Perhaps the most obvious source for the discrepancy in the number of unfamiliar items marked by each panel is disparate levels of expertise across panels. If some panels are composed uniformly qualified judges while other panels contain judges with low levels of content mastery, the patterns observed across panels are easy to understand. For this hypothesis to be true, however, it would require that each of the least qualified judges were assigned to a single panel. Given that judges were assigned to panels more or less randomly, this disparity in content knowledge seems extremely unlikely. Alternatively, it is possible that one or two unqualified judges are reasonable for the large number of unfamiliar items within a panel. Again, this does not appear to be the case in this data. Instead the results for Step 1 show that nine of the ten judges in panel one marked at least five items as unfamiliar. Panel three, on the other hand, had only one judge mark at least five items and only two judges marked any items at all. These results suggest that the discrepancy cannot be attributed to the behavior of one or two judges. In fact, the systematic pattern in the prevalence of omitted items suggests that some interpersonal panel effect is a more likely culprit.

An alternative hypothesis is that some group level effect resulted in some panels applying different standards when identifying items as "unfamiliar." This effect may be the result of some minor variations in the instructions provided by the facilitator or the effect of discussions between judges. Although it is impossible to know the specifics of these interactions it is easy to imagine that one vocal judge could alter the behavior of the entire panel. For example when the facilitator informs that the judges are not necessarily expected to be familiar with all items, a vocal judge could have said something like "That's good news because I haven't studied this material in ten years." This comment, although fairly benign, may free judges to mark items as unfamiliar without concern that they will be appear under-qualified. This sort of group level effect may reasonably explain both the discrepancy across panels and the consistency within panels.

These results indicate that under some circumstances judges would make different decisions regarding which items to mark as unfamiliar. This issue suggests that there is a disconnect between items which are truly unfamiliar to content experts and the items judges mark as unfamiliar. Given the prestige associated with being asked to serve as a content expert and the natural inclination to be respected by our peers it seems likely that under most circumstances judges would be reluctant to mark items as unfamiliar. If this effect is sufficiently prevalent it may ultimately be responsible for the extremely small number of unfamiliar items observed in the majority of panels. This reluctance to admit to limitations in their content expertise would suggest the number of marked items identified in each panel does not represent a true picture of the judges' content mastery. Instead this value can be thought of as a floor or lower bound of items which are truly unfamiliar. This would suggest that the change in Step 1 passing scores for panel one, may be an accurate representation of the true effect of unfamiliar items. According to this logic, the results observed in panels two and three may simply be a product of judges' reluctance to admit to gaps in their content mastery. Although this hypothesis cannot be demonstrated empirically, the large discrepancy in number of omitted items and the social pressure placed on content experts would seem to suggest that the effect of unfamiliar items presented in this study may represent a significant underestimation.

The results of this study indicate that unfamiliar items can have a deleterious effect on the validity of passing scores. When content experts are required to provide ratings to large numbers of unfamiliar items the resulting passing score may be significantly lower than the judge intended. This finding may not be generalizable to all Angoff standard setting scenarios, but it is a reasonable concern for practitioners working on any test where portions of the content may be unfamiliar to the content experts. Typically this concern may be greatest for tests which cover a broad range of highly technical content, but misalignment between content and judges has the potential to occur at any level. For example, K-12 testing programs like National Assessment of Educational Progress require that members of the public be included in the standard setting panels. Although these members are not selected at random, and are typically successful professionals, these judges may be many years removed from significant work with the tested content. Although the inclusion of these different stakeholders may be appropriate, these results suggest that these non-expert constituencies have the potential to dramatically suppress the recommended passing score.

The issue presented in this study is ultimately one of defining what it means for a person to be considered a content expert. Although some work has been done on the differences between experts and non-experts, this study represents the first empirical examination of how lack of content expertise at the item level affects passing scores. The results seem to suggest that even eminently qualified panelists cannot be expected to fully understand all content. Furthermore when unfamiliar items are identified the data in this study suggests that judges have a meaningful tendency to overestimate the difficulty of the items. When judges are required to rate all items, as in the common Angoff method, these lower ratings will tend to lower the passing score. These artificially low passing scores necessarily inflate the passing rate and have the potential in this case to license under qualified physicians. Although the magnitude of this problem was fairly modest in this study, the effect on passing rates is dependent on where the passing score is set relative to the distribution of examinees. Under the right circumstances the magnitude of errors seen in this study could increase the passing rate by more than 10% and place the public at significant risk.

This study builds a strong case for allowing and even encouraging content experts to omit unfamiliar items. Unfortunately with the common Angoff method there is no psychometrically sound procedure for setting standards on a subset of test items. An IRT-based Angoff method, on the other hand, allows for the estimation of recommended passing scores with only a subset of the total item pool. This flexibility does not ensure that judges will choose to omit all items outside the area of expertise but it does ensure that judges will not feel obliged to rate unfamiliar items. Eliminating this requirement has the potential to reduce a significant source of error and ultimately increase the validity of the recommended passing scores.

6.3 Adaptive Standard Setting

One of the most significant advantages of item response theory over classical test theory is that examinee ability and item difficulty can be placed onto a single scale. This feature allows items to be selected to maximize information for specific levels of examinee ability along the IRT proficiency scale. Typically for fixed form tests items are selected to maximize information in the area of the passing score, but at times items are selected dynamically to maximize information for each examinee. This adaptive approach offers significant advantages for both administration time and measurement error and has been used to great effect in many high profile testing programs including the GRE, GMAT, and NCLEX. Despite this success and popularity, adaptive algorithms have never been applied to standard setting. One potential advantage of the IRT Angoff method is the ability to adaptively select items for each judge to rate based on our current estimate of his or her conception about the performance level of the MCE. This study was the first to empirically test the accuracy and efficiency of adaptively set Angoff passing scores.

6.3.1 Summary of Results

The goal of this study was to determine if passing scores set using an adaptive standard setting technique were comparable to those set using traditional methods. To determine if an adaptive passing score could be considered comparable a distribution of acceptable passing scores was constructed using random samples of 150 items. The results of this process show that in order to be considered comparable, adaptively set passing scores need to fall into an extremely narrow range on the IRT proficiency scale. For Step 1 passing scores were required to fall into a range of 0.028 to be considered comparable for panel 3. Only slightly wider ranges were considered comparable for panels 1 and 2. For Step 2 the range of acceptable passing scores was slightly wider but still never exceeded 0.134. These tight tolerances are extremely important since they helped to ensure that adaptive standard setting would produce similar cut scores and passing rates to the traditional fixed form Angoff method.

The results of the simulated adaptive standard setting indicate that, despite the stringent criterion, three of the six panels produced comparable passing scores with 65 or fewer items administered. These results suggest that in principle an adaptive standard setting procedure can produce comparable passing scores to traditional standard setting methods. Unfortunately these results alone cannot be considered conclusive. Although three panels produced comparable results, three others did not, even when 75 items were administered. Although in two of these cases the observed passing score fell less than 0.05 outside the acceptable range, in one case the passing score fell more the 0.15 beyond the bounds of this distribution. These incongruent results seem to suggest that the adaptive algorithm applied in this study cannot be relied on to consistently produce standard setting results which are comparable to the tradition fixed form Angoff method.

The simulated results fail to provide clear evidence for the efficiency offered by the adaptive standard setting procedure. At times the adaptive method seems to offer comparable passing scores with significantly fewer items; at other times the method results in dramatically disparate passing scores. To illustrate the true benefits of an adaptive standard setting procedure, it is critical to understand the probability of observing these results when items are selected at random. If items selected at random provide meaningfully less accurate passing scores it would suggest that the adaptive method is beneficial for reducing administration time. Alternatively if passing scores established by selecting items at random are more accurate it suggests that the adaptive method offers little or no benefit. The results of this analysis indicate that for all panels when 65 or 75 items are used the adaptive results are not significantly better than what would be expected as a result of chance. Even when fewer items were used only one panel (Step 1 Panel 3) produced results that were significantly better than would be expected when items were selected at random. These results provide fairly clear evidence that the adaptive standard setting algorithm used in this study does not represent a significant improvement over a standard set with a random selection of test items.

The results of this study provide fairly clear evidence that the adaptive standard setting algorithm does not provide panels with a more efficient method of arriving at the same passing score. It is interesting to note, however, the degree to which a panel's recommended passing score remained consistent across the seven test length conditions. For example for two of the panels the passing score based on 15 items was within 0.001 of the recommended passing score based on 75 items. Although this level of reliability is not true for all panels, for five of the six panels results based on 45 items or more were within 0.100 of the passing score based on 75 items. These results suggest that although the adaptive algorithm may not consistently hone in on the “True” passing score for the complete item bank, the results are consistently driven to some other point on the IRT proficiency scale. This behavior would seem to suggest that the passing scores which judges recommend for the items with the most information may be systematically different from the passing score based on the items with the least information.

The systematic pattern in the passing scores observed in these results may suggest that adaptive algorithms which maximize information may be appropriate for estimating an examinee’s ability but may not be ideal in an adaptive standard setting procedure. Part of this issue may be directly linked to the purpose of high test information. Items which provide the most information tend to be highly discriminating. These highly discriminating items effectively separate examinees who have mastered the items content from those who have not, based on a dichotomous (right/wrong) decision. In a standard setting application, however, we are not forced to infer a position on the IRT proficiency scale based on a dichotomous decision. Instead Angoff probability estimates allow for deterministic identification of precise locations on the IRT scale. Highly discriminating items will still offer the greatest precision for specific parts of the IRT scale, but this percipience will be of little benefit if the judge's opinion about the ability of the MCE falls in a meaningfully different part of the IRT scale. In some circumstances this problem could become so severe that it would interfere with a judge's ability to express his or her true opinion.

Unfortunately the systematic patterns observed in the adaptively set passing scores would seem to suggest that the selection of the items with the greatest information is impeding judges' ability to express their view of the MCE's ability. These findings may be at least partially the result of an inter-correlation between the items' difficulty and discrimination parameters. In maximizing information the adaptive algorithm disproportionately selects the items with the highest a-parameter. If a- and b-parameters are correlated this interdependence would result in a majority of ratings coming from items in a particular part of the IRT proficiency scale. This clustering of highly discriminating items may tend motivate the standard to a specific point on the IRT scale, since a large portion of the 0-100 scale is devoted to a small range on the IRT scale. If judges believe the correct passing score is meaningfully outside this range the 0-100 ratings scale necessarily lacks precision in that portion of the IRT proficiency scale. This truncated scale combine with even modest levels of error may result in the majority of item level theta estimates being restricted to a narrow portion of the IRT scale. This tendency may help to explain both the consistency and the lack of accuracy observed in the passing score set using the adaptive algorithm.

6.3.2 Conclusion

The results of this study indicate that when standard setting items are adaptively selected to maximize information the resulting passing scores will often deviate meaningfully from the passing score based on the complete item bank. Although these results are discouraging, it is not clear to what degree these results would generalize to other testing contexts. It may be the case that the correlation between difficulty and discrimination parameters seen in this test would not be observed in other testing contexts. If highly discriminating items were equally distributed throughout the IRT scale an algorithm based on maximizing information may be able to supply greater precision without biasing the passing score. Alternatively it is possible that algorithms designed to optimize some other criteria like distance from the b-parameter may be appropriate in some circumstances. Based on results from this study, however, it is unreasonable to suggest that a similar adaptive standard setting method be implemented in practice. Further research is required to determine if these results can be remedied in other testing contexts.

Despite the errors introduced through the adaptive standard setting procedure, this study does suggest that reasonable passing scores may be obtainable with a dramatic reduction in the number of items administered. Although not specifically the focus of this study the consistency observed in the distribution of comparable passing scores provides some evidence that consistent passing scores could be obtained with a subset of the total item bank. Although further targeted research into this topic is certainly warranted, this study provides some evidence that the IRT-based Angoff method may deliver on the goal of accurate passing scores with a reduction in administration time.

6.4 Stability of Performance Standards

The final one of three studies examined the effect of systematic and random error on the estimation of passing scores. Random errors are introduced when judges struggle to produce internally consistent estimates of examinee performance across items. It was theorized that systematic errors would arise when the raw passing scores are mapped into the IRT proficiency scale. This study simulated standard setting results so that the magnitude of each of these sources of error could be compared across the common Angoff method and the IRT Angoff method. The purpose of this analysis was to determine if the use of an IRT-based Angoff method can improve the stability and accuracy of passing scores typically obtained by the True Score Angoff method. This study was the first to systematically evaluate the measurement properties of an IRT-based Angoff procedure.

6.4.1 Summary of Results

This study included three sets of analyses focused on stability and accuracy. The first set of analyses examined the role of random error in the estimation of passing scores. In this analysis the stability of recommended passing scores were compared across different selections of judges, ratings, and test items. The stability of the standard setting methods was calculated by comparing the standard deviations of the distribution of recommended passing scores across replications. The results indicated that recommended passing score for the mean and median IRT Angoff method varied with an average standard deviation across conditions of 0.279 and 0.293, respectively. These results were somewhat less stable than the average standard deviation observed across conditions for the True Score method of 0.223. These results suggest that even when using the exact same judges and ratings the True Score Angoff method produces noticeably more stable passing scores than either IRT based approach.

The second set of analyses examined potential systematic errors introduced into the passing score as a result of item difficulty. This analysis was based on the idea that the mapping of recommended test scores through the test characteristic curve onto the IRT proficiency scale may bias the resulting passing scores. To understand this effect, standard setting results were simulated for seven different tests with mean item difficulty ranging from -3.0 to 3.0. The results across the two IRT-based methods consistently produced passing scores close to the true passing score of 0.0. Deviations from this true value were random across the difficulty conditions and never exceeded 0.020. The results of the True

Score Angoff method, on the other hand, showed large biases introduced as a result of item difficulty. Across the seven conditions a deviation of 1.0 from the true passing score would result in a bias in the same direction of approximately 0.3. For example, if the mean item difficulty was two points lower than the judges' true opinion of the MCE's ability on the IRT scale the recommended passing score would be approximately 0.6 lower than the judges had intended. These results strongly suggest that passing scores based on the common Angoff method have the potential to be artificially bias as a result of the raw to scaled score transformation.

The final set of analyses in this study combined the effects of random and systematic error to understand the relative accuracy of IRT and True Score based passing scores. These two sources of error were evaluated by comparing the mean absolute difference between the true and observed passing score within each difficulty condition. The results show that across two IRT methods the mean absolute difference was fairly consistent. These errors are fairly modest and presumably can be attributed to random rather than systematic error. The results of the True Score Angoff method, on the other hand, show substantial errors systematically increasing as item difficulty deviates from judges' true opinion. These errors are a combination of random and systematic error which combine to produce fairly sizable errors across all difficulty conditions. The one notable exception is when the mean item difficulty is 0.0 and therefore is equal to the judges' true opinion there is no systematic error. Therefore under this condition the total error seen in the True Score method is lower than the error in either IRT Angoff method. On balance, however, the passing scores calculated using the IRT Angoff procedure tend to be less error prone than their True Score counterparts.

6.4.2 Conclusions

Although both random and systematic errors are significant sources of concern in standard setting, random errors are much easier to address. For the analysis conducted in this study, random samples of ten judges provided ratings 75 items. The random errors observed under these conditions can readily be moderated by increasing the number of judges or items. Furthermore since all recommended passing scores are based on a single panel, random errors could be further reduced by replicating results across panel. Therefore, even in the event of relatively large random errors, reasonable steps could be taken to increase the reliability of the recommended passing score.

Unfortunately, systematic errors like the ones seen in these True Score Angoff method results cannot be mitigated through a more complete sampling procedure. The results from this study suggest that at times the True Score Angoff passing score could be systematically biased as a result of specific test items. This is obviously a matter of concern when the passing score will be applied to multiple test forms over time, but the extent of the problem is not limited to tests with multiple forms. The fundamental issue is not that different forms will produce different passing scores. Rather the concern is that passing scores set on any particular form will not reflect the judges' true opinion because of the bias introduced by particular items on the test form. Although it is true that this bias will be eliminated if item difficulty is well aligned with the judges' underlying opinion, there is no way to assure this alignment in practice. Tests can be designed to offer information in the area of an existing passing score, but clearly they cannot be designed for passing scores which are yet to be set. These systematic errors potentially represent a serious threat to the validity of the recommended passing score. Since these errors cannot be resolved through additional sampling of items or judges, the results strongly suggest that these errors cannot be resolved within the confines of the True Score Angoff method.

Overall the results of this study suggest that the process of translating the recommended passing score onto the IRT proficiency scale has the potential to introduce serious and systematic errors into the passing score. Due to the systematic nature of these errors the resulting bias cannot be eliminated by sampling larger numbers of judges or items. Although this practice will increase the reliability of the passing score, it does nothing to ensure that the passing score accurately reflects the judges' opinions. Ultimately these results cast doubt on the validity of passing scores which undergo this raw to scaled score transformation. One potential solution to this issue is the use of the IRT Angoff procedure. Since the IRT-based approach places all rating onto the IRT proficiency scale before integrating the ratings across judges and items, the recommended passing score is not affected by the non-linear transformation. These results suggest that an IRT-based Angoff method would eliminated the systematic error introduced by item difficulty and increase the overall validity of the resulting passing scores.

6.5 Overall Discussion

The goal of the Angoff procedure is to infer the opinions of content experts and represent that opinion as a point along the score scale. Unfortunately the commonly applied Angoff method anchored in classical test theory has several limitations which interfere with its ability to properly estimate these opinions. Three of these limitations have been discussed in this thesis. First judges struggle to produce internally consistent ratings. Second, recommended passing scores are item dependent due to the non-linear raw to scaled score transformation. Finally, test, rather than item, level measurement requires that all judges provide ratings for all items. These limitations have the potential to introduce error into the standard setting process and may result in passing scores which fail to reflect the true opinions of the content experts. The IRT Angoff method was designed to mitigate these limitations and provide more valid and reliable passing scores. The three studies presented above, attempt to provide a thorough examination of the measurement properties of an IRT-based Angoff method. Together these results provide compelling evidence for the benefits of conceptualizing the Angoff method within an item response theory framework.

The logic undergirding the Angoff method is both appealing and straightforward. Judges internalize the ability of the minimally competent examinee and then estimate the proportion of MCEs who would answer each item correctly. The sum of these ratings for each judge is the recommended passing score on the raw score scale. Unfortunately despite the theoretical appeal of the procedure, content experts often times struggle to make the required judgments. Specifically, judges struggle to produce internally consistent estimates of examinee performance. Since these ratings are the mechanism through which the judge's expert opinion is inferred, internally inconsistent ratings fail to point to a single passing score and obscure the judge's true opinion. The IRT Angoff procedure was designed to mitigate this problem by pooling results across items and judges and using the median of the complete distribution of ratings. This approach was not designed to improve the consistency of individual judges, but instead to provide more reliable recommended passing scores for the panel. The results presented in chapter five provide clear evidence that the IRT-based Angoff method does not provide more reliable recommended passing scores. Based on these results it is reasonable to conclude that the IRT Angoff method fails to mitigate the unreliability introduced by judges' internally inconsistent ratings.

One of the requirements for the Angoff method is that each judge internalizes an ability associated with the minimally competent examinee. This ability can vary across judges and rounds but is expected to be consistent within each judge for a single round of judgments. Although these conceptions of examinee ability are thought to be invariant to item selection, the specific scale transformation used in the common Angoff method does not ensure score invariance across items. Although the magnitude of this issue had not been previously examined, the scale transformation posed a credible threat to the validity of the recommended passing scores. The IRT Angoff method addressed this issue by eliminating the need for a non-linear raw to scaled score transformation. Instead all ratings were immediately placed on the IRT scale prior to distilling them into the panel's recommended passing score. The results presented in chapter five clearly illustrate the benefit of this approach. In this study the score transformation was shown to result in significant systematic errors in the recommended passing score. Moving to the IRT-based standard setting approach eliminated these problems and resulted in accurate and systematically unbiased estimates of the recommended passing score. These results suggest that the IRT Angoff method effectively eliminates this limitation of the common Angoff method.

Because the common Angoff method is grounded in classical test theory the standard setting results provide test rather than item level measurement. This means that passing scores set using the True Score Angoff method require that all judges rate all test items. This requirement is not only inefficient but has the potential to bias the recommended passing score. Chapters three and four examined how the item level measurement offered by the IRT Angoff method could provide more efficient and accurate passing scores. Chapter four examined the degree to which items could be adaptively selected to produce comparable passing scores with less administration time. The results of this analysis showed that although the specific adaptive algorithm used in this analysis did not produce comparable passing scores, there is some evidence to suggest that reasonable passing scores could be obtained with a random subset of test items. Chapter three examined the extent to which test level measurement has the potential to artificially bias passing scores. The results indicate that judges do provide systematically bias rating to unfamiliar items. Under some circumstances, depending on the number of unfamiliar items, these systematic errors have the potential to significantly suppress passing scores. These results suggest that the item level measurement provided by the IRT Angoff method may offer a significantly more valid passing score, while potentially offering some additional efficiency.

The purpose of this study was to examine the theoretical benefits of interpreting Angoff ratings within an item response theory framework. Although this was not the first study to use IRT for the interpretation of Angoff standard setting results, this study did represent the first comprehensive analysis of the measurement properties of Angoff passing scores set within a modern test theory framework. The results presented in chapters three four, and five deliver on this promise and provide considerable insight into the benefits of an IRT-based Angoff method. The results suggest that interpreting Angoff standard setting results within an IRT framework offers a number of significant advantages over the more common True Score Angoff method. These advantages include the mitigation of two potentially significant sources of systematic error which would improve the validity of recommended passing scores. Although the analyses presented in this study failed to empirically confirm all of the theoretical advantages of the IRT Angoff method, the findings provide strong evidence in favor of an IRT-based approach to the Angoff method. Overall these results have important implications for how passing scores are set and evaluated. This research could lead to more accurate passing scores and ultimately more valid high stakes decisions.

6.6 Limitations

Although this study has important implications for standard setting several limitations must be considered when interpreting these results. One significant limitation is the generalizability of results based on medical licensing data. Although we were extremely fortunate to have access to a high quality pool of operational standard setting data, features of the data which are unique to the USMLE may have limited the generalizability of these results in other testing contexts. For example because the population of medical school graduates both extremely capable and fairly homogeneous the IRT item parameters are dissimilar to those seen in most K-12 testing contexts. Furthermore, the USMLE covers an exceptionally wide domain of content. This means that the content experts serving on USMLE standard setting committees may be less familiar with any given item then content experts working in other, more narrowly defined, content domains. These results may reasonably be expected to generalize beyond medical licensing to other high stakes credentialing exams; however, generalizations beyond these contexts to K-12 testing may not be justified.

In addition to concerns regarding the generalizability of the findings, the interpretation of the selective standard setting results may be limited by the identification of unfamiliar items. For the selective standard setting analysis unfamiliar items were self-identified by individual judges. Although the data set was reasonable for our analysis, this approach was limited by differences in judges' propensity to recognize or concede that item content was unfamiliar. This limitation does not suggest that the conclusions regarding that analysis are incorrect, but it does make it difficult to understand the complete scope of the problem. Given this limitation it is difficult to separate the objectively unfamiliar items from those which are marked as unfamiliar.

Finally this study is potentially limited by lack of IRT model fit. In item response theory, many of the desirable measurement properties including invariance and item level measurement are dependent on good model fit. In this study model fit was not explicitly evaluated. If violated this model fit assumption may meaningfully impact the results. The decision to forgo explicit examination of model fit was based on the idea that all analyses were conducted using standard setting results from an operational testing program. It was therefore assumed that model fit had been evaluated as part of the test development process. Although a violation of model fit would have a deleterious effect on the results of this study, the impact would be far more consequential for the test development and scoring procedures. Since it is only recommended that the IRT Angoff method be applied to operational testing programs using an IRT model which fits the data, model fit analysis should be conducted prior to implementing IRT in any operational activity.

6.7 Future Research

The findings from this research suggest that the IRT Angoff method require future research both to address the limitations of this study, and to expand and clarify its conclusions. From the perspective of practitioners perhaps the most important limitation is the lack of evidence demonstrating the applicability of the IRT Angoff method to other testing contexts such as K-12 achievement testing. Increasingly categorical decisions based on these achievement tests have high stakes implications for schools, teachers and students. Although the specifics of the standard setting method vary across states, the findings from this study suggest that the validity of the passing score may be in question when the common True Score Angoff method is employed. The IRT Angoff method could potentially offer significant improvements to the validity of passing scores by eliminating several sources of systematic error. Future research should be devoted to the evaluation of the appropriateness of employing the IRT Angoff method within a K-12 testing environment.

Future research is needed to address the use of self-identified unfamiliar items in the selective standard setting analysis. Difference in the judges' willingness to mark items as unfamiliar has the potential to significantly impact the perceived magnitude of the error introduced by unfamilar items. This suggests that an objective measure of item familiarity and comfort may be critical to better understanding the full extent of these errors. Although no pure objective measure of familiarity is available, it may be reasonable to use content mastery as an acceptable proxy. In this context judges could be asked to provide answers to each test item prior to making their Angoff judgment. Although it is fairly common to ask judges to answer test items during the training process, this would be different in that responses would be collected by the facilitator. When judges answer the item incorrectly these items would be considered unfamiliar, while correct answers would be considered familiar. A future study could replicate the selective standard setting analysis using this new method for flagging unfamiliar items. This study would provide an empirical objective measure of judge familiarity, or at least content mastery, for each item.

In addition to future research to address limitations, this study has introduced valuable concepts which could reasonably be used in a variety of future research designed to improve the validity of passing scores. One of the key advantages of the IRT Angoff method is that it produces a distribution of ratings for each judge. This distribution provides an empirical method to compare the internal consistency of ratings across judges. Since internal consistency is a critical source of validity evidence for passing scores, the IRT Angoff approach could be used to study the impact of different interventions on judge's internal consistency. For example research could be conducted on the effect of training on judges' internal consistency. Additionally the method could be used in the development of new training tasks geared specifically to improving judges' understanding of which items are empirically difficult and which are empirically easy. Although training is an obvious avenue of future research, the IRT Angoff method could also be used to facilitate the study of the effect of specific types of discussion or performance data on judge's ratings. By providing an empirical method for evaluating the internal consistency of judges' ratings the IRT Angoff method could support a broad variety of research into the benefits of different standard setting interventions.

In addition to supporting research on standard setting procedures, the IRT Angoff method could be used in future research to evaluate both judges and items. Because the IRT Angoff method provides an objective measure of a judge's internal consistency, the method would allow for a consistency criterion to be established prior to the standard setting meeting. This criterion could be based on a variety of factors but would presumably be grounded in the desired standard error around the panel's recommended passing score. Judges could then be removed from the panel entirely or required to undergo additional training until their internal consistency had met a predetermined threshold. In addition to identifying judges who struggle to produce internally consistent passing score, the IRT Angoff method could be used to identify items which elicit aberrant ratings from content experts. This could be achieved by analyzing item characteristics of items which tend to appear in the tails of each judge's distribution of ratings. Although it will typically not be appropriate to remove these items, it may be appropriate to provide specific training or devote specific discussion to these item types. By identifying and ameliorating the impact of inconsistent judges and items the IRT Angoff method could facilitate valuable future research to improve the overall validity of passing scores.

REFERENCES

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.; pp. 508-600). Washington, DC: American Council on Education.

Bourque, M.L., & S. Byrd. (Eds.). (2000). *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements*. Washington, DC: National Assessment Governing Board.

Busch, J.C., & Jaeger, R.M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27,* 145-163.

Chang, L., Dziuban, C.D., Haynes, M.C., & Olson, A.H. (1996). Does a standard reflect minimal competency of examinees of judge competency? *Applied Measurement in Education, 9(2), 161-173.*

Cizek, G. j. (Ed.). (2001). *Standard setting: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Clauser, B.E., Harik, P., Margolis, M.J., McManus, I.C., Mollon, J., Chis, L., & Williams, S. (2009). Empirical evidence for the evaluation of performance standards estimated using the Angoff procedure. *Applied Measurement in Education, 22,* 1-21.

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J., & Dillon, G.F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement,* 46(4), 390-407.

Clauser, B. E., Mee, J., & Margolis, M. J. (2011, April). The effect of data format on integration of performance data into Angoff judgments. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Clauser, B.E., Swanson, D.B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39,* 269-290.

Clauser, J.C., Clauser, B.E., & Hambleton, R.K. (2011, April). *Variability across judges in their ability to estimate probabilities: The case for weighting.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Clauser, J.C., Sireci, S.G., & Clauser, B.E. (2010, October). *The effect of performance data on the validity of standard setting*. Paper presented at the meeting of the Northeast Educational Research Association Conference, Rocky Hill, CT.

Davey, T., Fan, M., & Reckase, M.D. (1996). *Some new methods for mapping ratings to theNAEP θ-scale to support estimation of NAEP achievement level boundaries*. Paper presented at the meeting of the National Council on Measurement in Education, New York.

Ferdous, A. A., & Plake, B. S. (2005). The use of subsets of test questions in an Angoff standard-setting method. *Educational and Psychological Measurement*, 65 (2), 185-201.

Ferdous, A. A., & Plake, B. S. (2008). Item response theory-based approaches for computing minimum passing scores from Angoff-based standard-setting study*. Educational and Psychological Measurement. 68* (5),778-796.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18*, 223-232.

Goodwin, L.D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline candidates. *Applied Measurement in Education, 12(1)*, 13-28.

Gravetter, F. J., & Wallau, L. B. (2009). *Statistics for the behavioral sciences* (8th ed.). United States: Wadsworth Cengage Leaning.

Gross, M.E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation and the Health Professions, 9*, 267-285.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*, 355–366.

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

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–55.

Hambleton, R. K., & Swaminathan H. (1986). *Item response theory: Principles and applications.* Boston, MA: Kluwer Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications, Inc.

Hays, W. L. (1994). Statistics (5th ed.). Fort Worth, United States: Harcourt Brace College Publishers.

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

Kane, M.T. (1994). Validating the 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. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Earlbaum Associates, Publishers.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

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

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

Meara, K., 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.

Mee, J., Clauser, B. E., & Margolis, M. J. (2011, April). *The impact of process instructions on judges' use of examinee performance data*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

National Research Council. (1999). Setting reasonable and useful performance standards. In J. W. Pelligrino, L. R. Jones, & K. J. Mitchell (Eds.), *Grading the nation's report card:*

Evaluating NAEP and transforming the assessment of educational progress (pp. 162–184). Washington, DC: National Academy Press.

Nering, M.L. & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Plake, B. S., & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment*, 7, 87-97.

Plake, B.S., Impara, J.C., & Irwin, P. (1999, April). *Validation of Angoff-based predictions of item performance*. Paper presented at the meeting of the American Educational Association, Montreal, Quebec, Canada.

Plake, B.S., Impara, J.C., & Irwin, P.M. (2000). Consistency of Angoff-based predictions of item performance: Evidence of the technical quality of results from the Angoff standard setting method. *Journal of Educational Measurement, 37*, 347-356.

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.

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

Reckase, M. D. (2000). A survey and evaluation of recently developed procedures for setting standards on educational tests. In M.L.Bourque (Ed.), *Setting performance standards on the national assessment of educational progress: Affirmation and improvements* (pp. 41-70). Washington D.C.: National Assessment Governing Board.

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. 159-174). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Saunders, J.C., Ryan, J.P., & Huynh, H. (1981). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. *Applied Psychological Measurement, 5*, 209-217.

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

Sireci, S.G., & Clauser, B.E. (2001). Practical issues in setting standards on computerized adaptive tests. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 355-370). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Sireci, S. G., Patelis, T., Rizavi, S., Dillingham, A. M., & Rodriguez, G. (2000, April). *Setting standards on a computerized-adaptive placement examination.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Skorupski,W. P., & Hambleton, R. K., (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18(3)*, 233-256.

Smith, J.E., (1999). *Using IRT created models of ability in standard setting* (Unpublished doctoral dissertation) University of Nebraska, Lincoln.

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

Swanson, D. B., Dillon, G. F., & Ross, L. E. P. (1990). Setting content-based standards for national board exams: initial research for the Comprehensive Part I Examination. *Academic Medicine*, 65 (9, Suppl.) S17-S18.

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

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R.D. (2003). BILOG-MG (Version 3) [computer software]. Lincolnwood, IL: Scientific Software International, Inc.