

2
2010

LIBRARY
Michigan State
University

This is to certify that the
dissertation entitled

A COMPREHENSIVE ITEM RESPONSE THEORY FRAMEWORK
FOR EVALUATING STANDARD SETTING

presented by

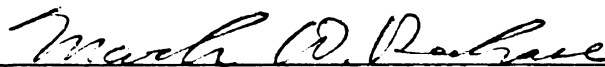
Adam E. Wyse

has been accepted towards fulfillment
of the requirements for the

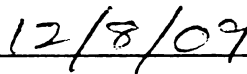
Ph.D.

degree in

Measurement & Quantitative
Methods



Major Professor's Signature



Date

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE
SEP 19 2011		

**A COMPREHENSIVE ITEM RESPONSE THEORY FRAMEWORK
FOR EVALUATING STANDARD SETTING**

By

Adam E. Wyse

A DISSERTATION

**Submitted to
Michigan State University
in partial fulfillment
for the degree of**

DOCTOR OF PHILOSOPHY

Measurement & Quantitative Methods

2009

ABSTRACT

A COMPREHENSIVE ITEM RESPONSE THEORY FRAMEWORK FOR EVALUATING STANDARD SETTING

By

Adam E. Wyse

The last few decades have seen the increased use of standard setting procedures to set cut scores on educational and psychological assessments. These cut scores are used for classifying students into different performance categories and for making high stakes decisions in educational, psychological, professional licensure and certification testing situations. The fundamental assumption behind the use of cut scores is that they represent the achievement levels educators, policy makers and stakeholders intended when the performance standards were formulated. That is, cut scores are assumed to be unbiased and precise representations of the cut scores that panelists had in mind when they set them. Although researchers recognize the importance of these properties, few procedures exist for determining whether cut scores are unbiased.

This study proposes a comprehensive item response theory (IRT) framework for evaluating cut scores established through a standard setting process. This new framework includes a step-by-step process for evaluating cut scores from any IRT-based standard setting procedure in simulated or operational situations when assessment data can be assumed to fit an IRT model. Specifically, this framework extends Reckase's (2006a) psychometric theory for standard setting, which assumes that an individual panelist has a hypothetical cut score that they intend to set when providing standard setting judgments. Construct maps (Wilson, 2005) aid Reckase's (2006a) psychometric theory and are used

to provide a spatial representation of the relationship between the score scale underlying the assessment and examinee and item statistics derived from an IRT model.

Examples of how this new framework can be used to formulate indices to evaluate cut scores established from the Angoff method with Mean Estimation and a version of the Bookmark method known as Mapmark on the National Assessment of Educational Progress (NAEP) are provided. Results suggest that cut score biases and inconsistencies could impact individual panelist and group cut scores when both the Mapmark and Angoff procedures are used. An important finding is that cut score biases appear to be more of a concern in earlier rounds of standard setting and at the basic cut score. Investigations of the impact of biases on the percentage of students above the cut score suggest that biases for individual panelists could have a large impact on the percentage of students above the cut score for them. The group panelist biases do not appear to have a large impact on the percentage of students classified as being above the cut score except for the basic cut score with Angoff procedure and the basic and proficient cut scores with the Mapmark procedure.

An important outgrowth of the new framework is an explicit recognition that there are two potential issues that can produce bias in cut score estimates. These two potential issues include (1) the possibility for gaps in the score scale from lack of standard setting stimuli at every score scale location and (2) the possibility for rater inconsistency. These two issues may also work in concert to produce bias in cut scores. Important distinctions are also made between what it means to evaluate cut scores and what it means to evaluate standards. Finally, some limitations of the new framework as well as some areas for future research are also identified.

Copyright by

Adam E. Wyse

2009

DEDICATION

This dissertation is dedicated to my God and Savior. It is through Him that all wisdom, knowledge, and understanding are possible. I also dedicate this dissertation to my wife and family who have continually supported and believed in me.

ACKNOWLEDGEMENTS

There are many people that have helped me tremendously as I have completed this dissertation. First, I would like to thank my advisor Dr. Mark Reckase who has provided me with great insight and support as I have pursued my degree. I could not have asked for a better advisor. I also thank my dissertation committee members Dr. Ryan Bowles, Dr. Sharif Shakrani, and Dr. Edward Roeber who have challenged and encouraged me along the way. I also thank Dr. Barbara Schneider and the Office of the Hannah Chair at Michigan State University who have supported my research and pushed me to become a better scholar and researcher throughout my years at Michigan State. I also deeply indebted to Susan Loomis from NAGB, Teri Fisher from ACT, and Dr. Matt Schulz from Pacific Metrics who talked to me on several occasions about my project and helped me to obtain data to complete my dissertation. Dr. George Engelhard and Dr. Michael Kane also shared insights with me that engaged me in this research area and motivated me to pursue this topic. I also have benefited greatly from many conversations with Venessa Keesler, Dr. Raymond Mapuranga, Dr. Joseph Martineau, Dr. Dipendra Subedi, and Steve Viger about my work while I was a graduate student. Dr. Raymond Mapuranga and Dr. Joseph Martineau both provided me with valuable feedback and comments on my dissertation at various stages along the way. Lastly, I thank the many other professors, graduate students, and friends that have enriched my time and studies at Michigan State including Tim Ford, Dr. Ken Frank, Dr. Cassie Guarino, Dr. Steve Haider, Nate Jones, Yun-Jia Lo, Dr. Kim Maier, Dr. Yeow Meng Thum, Dr. Tenko Raykov, Qiu Wang, Yisu Zhou, and numerous others.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	x
KEY TO ABBREVIATIONS	xi
CHAPTER 1 INTRODUCTION	1
1.1 Standard Setting Process.....	3
1.1.1 Step 1: Call for Standards and Definition of Policy	5
1.1.2 Step 2: Define Performance Level Descriptors	7
1.1.3 Step 3: Select a Standard Setting Method.....	9
1.1.4 Step 4: Choose a Standard Setting Panel and Design	9
1.1.5 Step 5: Train Panelists to Use the Standard Setting Method	11
1.1.6 Step 6: Collect Panelist Ratings.....	12
1.1.7 Step 7: Compile Panelist Ratings and Obtain Cut Score Estimates.....	13
1.1.8 Step 8: Provide Feedback and Facilitate Discussion	14
1.1.9 Step 9: Conduct Panelist Evaluations	14
1.1.10 Step 10: Prepare Technical Documentation and Validity Evidence.....	15
1.1.11 Step 11: Review Documentation and Determine Final Cut Scores	16
1.2 Angoff and Bookmark Methods	16
1.2.1 Angoff Method.....	17
1.2.2 Bookmark Method	18
1.2.3 Overview of Angoff Derivative Methods.....	19
1.3 Formulation of Problem.....	28
1.4 Purpose.....	30
CHAPTER 2 LITERATURE REVIEW	33
2.1 Kane's Validity Framework for Standard Setting	33
2.1.1 Procedural Evidence	34
2.1.2 Internal Evidence	36
2.1.3 External Evidence	37
2.1.4 Summary of Strengths and Weaknesses of Kane's Validity Framework	38
2.2 Engelhard's Rasch Evaluation Framework.....	39
2.3 Reckase's Psychometric Theory	43
2.3.1 IRT Models.....	43
2.3.2 Link between IRT and Reckase's Psychometric Theory	49
2.3.3 Previous research Using Reckase's Framework	51
2.4 Motivation for New Framework	52
2.5 Previous Comparisons of Angoff and Bookmark Methods.....	53

CHAPTER 3 NEW EVALUATION FRAMEWORK	58
3.1 Extensions of Reckase's Psychometric Framework	58
3.2 Construct Maps	61
3.3 New Comprehensive Evaluation Framework	65
3.3.1 Step 1: Create a Construct Map	67
3.3.2 Step 2: Examine the Construct Map and Determine Relationships	68
3.3.3 Step 3: Determine How the Method is to be Evaluated	70
3.3.4 Step 4A: Simulate the Standard Setting Method	71
3.3.5 Step 5A: Examine Recovery of Simulated Cut Scores	73
3.3.6 Step 4B: Develop a Statistical Model or Index to Evaluate Method	74
3.3.7 Step 5B: Evaluate Standard Setting Method	77
3.3.8 Step 6: Draw Conclusions and Make Recommendations	78
CHAPTER 4 INDICES FOR EVALUATING ANGOFF AND BOOKMARK	79
4.1 Data to Illustrate the New Evaluation Methods	80
4.2 Methods for Evaluating Angoff Method Outcomes	81
4.2.1 Indices for Evaluating Angoff Method Outcomes	85
4.3 Methods for Evaluating Bookmark Method Outcomes	88
4.3.1 Indices for Evaluating Bookmark Method Outcomes	91
CHAPTER 5 COMPARISON OF ANGOFF AND MAPMARK METHODS	96
5.1 Description of 2005 NAEP Mathematics Pilot Study Data and Procedures	96
5.2 Analysis Procedures	97
5.3 Results for the Angoff Method	102
5.4 Results for the Mapmark Method	114
5.5 Comparison of Potential Biases Between Angoff and Mapmark Methods	117
5.6 Practical Impact of Potential Biases	120
5.7 Discussion of Empirical Comparisons	131
CHAPTER 6 CONCLUSION	135
6.1 Unique Contributions	136
6.2 Limitations and Concerns	139
6.3 Future Research	143
REFERENCES	147

LIST OF TABLES

1.1	NAEP 2005 12 th Grade Mathematics Basic Performance Level Descriptor	8
1.2	Angoff Derivative Standard Setting Methods.....	20-25
3.1	Hypothetical Mathematics Construct Map	64
3.2	Construct Map for Hypothetical Booklet Standard Setting Method.....	68
3.3	Individual Panelist Ratings for Hypothetical Booklet Standard Setting.....	75
4.1	Item Parameters for Nine Items Used in the Didactic Examples.....	80
4.2	Construct Map for the Angoff Method	82
4.3	Example of a Hypothetically Inconsistent Rater	84
4.4	Construct Map for the Bookmark Method with an RP of 0.67.....	90
5.1	Potential Panelist Biases in Angoff Ratings	103-107
5.2	Panelist Inconsistencies in Angoff Ratings.....	108-112
5.3	Potential Group Biases for the Angoff Method	112
5.4	Potential Group Inconsistencies for the Angoff Method	113
5.5	Maximum Potential Panelist Biases for the Mapmark Method	115-116
5.6	Maximum Potential Group Biases for the Mapmark Method	116
5.7	Differences between Cut-Score Estimates and Biases.....	119
5.8	Changes in Angoff Method PAC for Individual Panelists.....	121-126
5.9	Changes in Angoff Method PAC for Groups of Panelists.....	127
5.10	Changes in Mapmark Method PAC for Individual Panelists.....	128-130
5.11	Changes in Mapmark Method PAC for Groups of Panelists.....	131

LIST OF FIGURES

1.1	Steps Involved in a Typical Standard Setting Process.....	5
2.1	Example of an Item Characteristic Curve for Two Items	47
2.2	Example of a Test Characteristic Curve for 10 Rasch Items	49
3.1	Framework for Evaluating a Standard Setting Procedure.....	66

KEY TO ABBREVIATIONS

AYP	Adequate Yearly Progress
CCSSO	Council of Chief State School Officers
GPCM	Generalized Partial Credit Model
ICC	Item Characteristic Curve
IRT	Item Response Theory
MCE	Minimally Competent Examinee
MRM	Multifaceted Rasch Model
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NCLB	No Child Left Behind
OIB	Ordered Item Booklet
PAC	Percent above Cut-score
PCM	Partial Credit Model
PLD	Performance Level Descriptor
RP	Response Probability
TCC	Test Characteristic Curve
2PL	Two-Parameter Logistic
3PL	Three-Parameter Logistic

CHAPTER 1

INTRODUCTION

Over the last few decades, the focus on and scrutiny of student educational achievement in K-12 settings has reached new heights. This change has resulted in an increased emphasis on educational accountability and equality of educational opportunity, which has had a resounding impact on the United States educational system (Koretz & Hamilton, 2006). Under the Bush administration, educational performance and accountability took center stage in the form of the No Child Left Behind Act (NCLB, 2001). NCLB emphasizes school accountability and educational equality by tracking school and student performance in each state on high stakes educational assessments (Linn, 2003a; Linn et al. 2003; Porter, et al., 2005). Inherent to the success of the new educational accountability systems are the standards and corresponding cut scores developed for measuring “continuous and substantial yearly improvement of each school and local education agency” (Goertz, 2001) –known as adequate yearly progress (AYP) (IASA, 1994).

An important component of the aforementioned educational accountability context is educational assessment *standards*. These are defined as achievement goals for examinees on an assessment which are set up to classify examinees into different levels of performance. In most cases, a standard is defined in relation to a performance level descriptor (PLD). PLDs are written descriptions of the knowledge, skills, and abilities that examinees at a particular level of test performance would be expected to possess if they are to be classified at that level of performance (Cizek & Bunch, 2007; Perie, 2008).

PLDs are typically operationalized through a process called standard setting where cut scores are derived by panels of stakeholders (Cizek & Bunch, 2007). In most cases, PLDs are defined by the organization responsible for setting standards prior to the use of a standard setting procedure (Cizek & Bunch, 2007; Perie, 2008). Usually, the organization defines several PLDs that correspond to several distinct levels of performance. For example, the National Assessment of Educational Progress (NAEP) — an assessment administered by the federal government to track student performance at fourth, eighth, and twelfth grade in mathematics, reading, writing, civics, history, geography, economics, arts, and science — has three PLDs: one for basic, one for proficient, and one for advanced levels of performance (Pellegrino, et al., 1999; Reckase, 2000; Loomis & Bourque, 2001). In other words, NAEP PLDs define three standards that separate students into four categories of test performance: below basic, basic, proficient, and advanced.

The terms “standard” and “cut score” are often used interchangeably, but do not mean the same thing. As noted previously, the term “standard” refers to achievement goals that are set up to categorize examinees into different levels of test performance and are articulated in terms of descriptions (i.e., PLDs) of what examinees should know and be able to do at each of level of performance. A cut score, on the other hand, is the location on a scoring continuum (e.g., a score scale) that is used to distinguish among examinees at different levels of performance. Therefore, the cut score is usually a single number on the score scale. Hence, one might view the cut score as the operational definition of the standard (Kane, 2001; Reckase, 2001). In summary, standard setting is

used for translating a standard into a cut score that can then be used to make classification decisions (e.g., pass/fail, proficient/not proficient) based on examinee test performance.

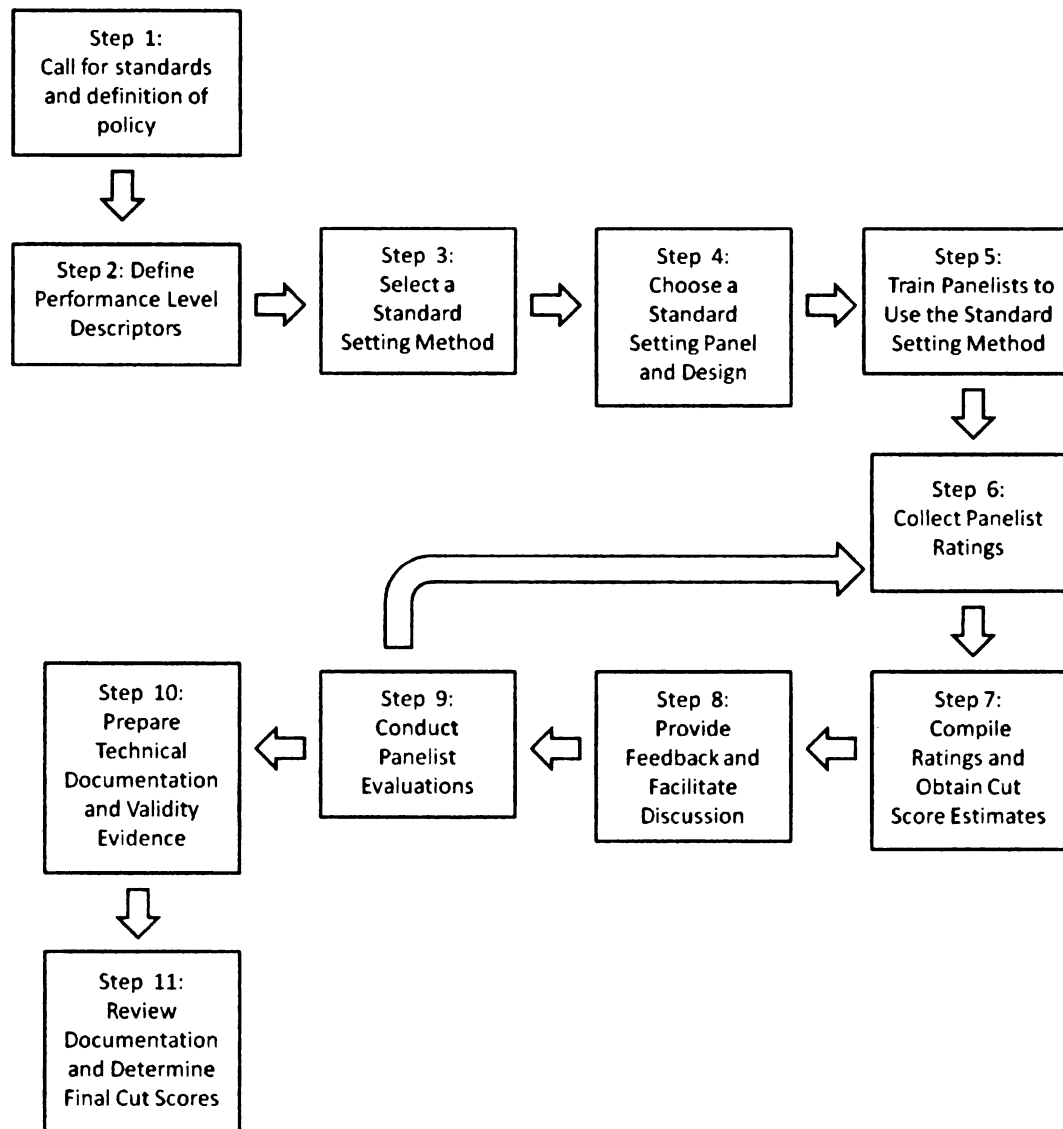
Broadly speaking, the term “standard setting” is a misnomer because people who participate in standard setting usually do not set standards. The standards are usually set in advance by policy boards that create the PLDs. The individuals who participate in standard setting are asked to interpret these definitions and translate the standard onto some scoring continuum in the form of a cut score. In this dissertation, the term standard setting refers to the process of translating standards into cut scores. An explanation of the standard setting process is provided in the next section.

1.1 Standard Setting Process

The process of developing standards and their associated cut scores on large scale assessments is a complicated and multi-step process. Depending on the selected procedure and the amount of information deemed necessary for creating accurate and unbiased cut scores, the process can range from five to as many as twelve steps. To ground the discussion in a common situation in which cut scores are typically derived, the steps involved in the NAEP standard setting processes and how NAEP completes each step are reviewed. NAEP was chosen as an example in this dissertation because it is a commonly used large scale assessment program used to make important educational policy recommendations and the standard setting processes used on NAEP have received considerable attention and scrutiny (Shepard, et al., 1993; Pellegrino, et al., 1999; Hambleton, et al., 2000).

The steps reviewed are consistent with the ones that were undertaken to establish cut scores in the standard setting processes that are used as examples in later sections of this dissertation. Some of these steps overlap with the suggestions of Reckase (2000) and Hambleton and Pitoniak (2006), while others are introduced to provide a clear picture of the whole standard setting process. A schematic of the eleven steps used in most NAEP standard settings is provided in Figure 1.1 and each step in the process flow will be described below.

Figure 1.1: Steps Involved in a Typical Standard Setting Process



1.1.1 *Call for Standards and Definition of Policy*

The first step in the standard setting processes is a call for standards and the definition of standard setting policy by the agency responsible for creating the standards. This first step in the process includes providing a rationale for the standards, as well as

how the standards will be used and interpreted (Reckase, 2000). Typically, the first step in the process also involves selecting the assessment instruments, describing how the standards will be reported, and characterizing who will be impacted by the standards as well as the stakes attached to the standards (Reckase, 2000).

In NAEP standard setting, the National Assessment Governing Board (NAGB) is responsible for creating the standards and defining standard setting policy. This includes overseeing the development of the assessment instruments, defining the content that is the focus of each assessment, and deciding on the score scales and number of achievement levels that will be used to report results. NAGB is also responsible for reporting results to the public after the assessment is administered.

As part of this first step in the process, NAGB also develops generic policy descriptions and labels for achievement levels (e.g. advanced, proficient, basic) that it applies to all content areas and grade levels. The label is a name for a level of performance, such as advanced, proficient, or basic, on an assessment. The basic level on NAEP is defined as follows:

Basic: This level denotes partial mastery of prerequisite knowledge and skills fundamental for proficient work at each grade. (Reckase, 2000, p. 6).

The generic policy descriptions of each level of achievement are later refined into more specific statements about what students should know and be able to do at each performance level on a particular assessment in the next step in the standard setting process.

1.1.2 Step 2: Define Performance Level Descriptors

The next step in most standard settings is the development of the PLDs by the policy board that defines standards for a particular assessment. The PLD is the written description of what it means to be classified at that particular level of performance on the test. PLDs are more specific than the generic policy descriptions that are developed in the first step of the process. They describe the specific knowledge, skills, and abilities that students at each achievement level on each particular assessment would be expected to have if they were classified into that performance level.

In NAEP standard setting, the NAGB brings together a panel of experts and stakeholders to define the PLDs. NAGB develops three standards and corresponding PLDs that are used to separate students into four levels of performance (below basic, basic, proficient, and advanced). An example of a PLD for the basic level of the 2005 mathematics standard setting is provided in Table 1.1.

The essential part of this step is making sure that the PLDs are clear and easy to understand because the panelists will be asked to translate them into cut scores. A detailed discussion of how to create PLDs in standard setting can be found in Perie (2008).

Table 1.1: NAEP 2005 12th Grade Mathematics Basic Performance Level Descriptor

BASIC

Twelfth-grade students performing at the basic level should be able to solve mathematical problems that require the direct application of concepts and procedures in familiar situations. For example, they should be able to perform computations with real numbers and estimate the results of numerical calculations. These students should also be able to estimate, calculate, and compare measures and identify and compare properties of two- and three-dimensional figures, and solve simple problems using two-dimensional coordinate geometry. At this level, students should be able to identify the source of bias in a sample and make inferences from sample results, calculate, interpret, and use measures of central tendency and compute simple probabilities. They should understand the use of variables, expressions, and equations to represent unknown quantities and relationships among unknown quantities. They should be able to solve problems involving linear relations using tables, graphs, or symbols; and solve linear equations involving one variable.

Number Properties and Operations:

- Perform computations with real numbers including common irrational numbers or the absolute value of numbers
- Solve problems involving factorization and divisibility
- Estimate the results of numerical calculations including square and cube roots of numbers, or very small and very large numbers

Measurement and Geometry:

- Recognize, define, and describe properties of two and three dimensional figures
- Estimate, calculate, and compare measures of two and three dimensional figures
- Draw or sketch a geometric figure from a description
- Use the Pythagorean Theorem to solve problems in two dimensions
- Solve problems in coordinate geometry (two dimensions)

Data Analysis and Probability:

- Evaluate a sample for bias and make inferences from sample results
- Describe the impact of outliers on measures of central tendency and variability
- Calculate, interpret, and use measures of central tendency and variability
- Understand the use of correlation coefficients to describe the relation between two data sets
- Compute simple probabilities
- Distinguish between experimental and theoretical probability

Algebra:

- Understand the use of variables, expressions, and equations to represent unknown quantities and relationships among unknown quantities
- Solve problems involving linear relations expressed in algebraic, verbal, tabular, or graphical forms
- Solve linear equations in one variable
- Perform basic operations on algebraic expressions
- Recognize, describe, and extend arithmetic or geometric progressions

1.1.3 Step 3: Select a Standard Setting Method

After the policy board develops the PLDs, the next step is to select a standard setting method. There are many possible standard setting methods that can be chosen. Most of these methods involve collecting judgments from panelists, the people who provide the ratings, about where they think the cut score should be set. Standard setting procedures are often separated into examinee-centered (procedures that focus on judgments related to examinees and their relationship to the PLD) or test-centered approaches (procedures that focus on test content and/or test items and their relationship to the PLD) (Jaeger, 1989, Kane, 1994; Berk, 1996; Cizek, 2001; Cizek & Bunch, 2007); although this classification scheme has become blurred in the recent literature on standard setting (Hambleton & Pitoniak, 2006). Most of the commonly used standard setting methods are described in an edited book by Cizek (2001), a book by Cizek and Bunch (2007), and a recent chapter by Hambleton and Pitoniak (2006).

NAEP has used both test-centered and examinee-centered approaches to set cut scores in different contexts. These methods include variations of the Angoff (Angoff, 1971) and Bookmark (Lewis, et al., 1996) standard setting methods that are empirically compared and described in greater detail in subsequent sections of this dissertation.

1.1.4 Step 4: Choose a Standard Setting Panel and Design

The fourth step involves recruiting panelists and deciding on a design that will be used to determine the cut scores. Panelists are typically recruited from a pool of representative stakeholders who would be influenced by the decisions made based on the cut scores. In NAEP standard setting, the standard setting panel is selected so that 70% of

the panel represents teachers and other educators and 30% of the panel is non-educators, such as community leaders, military personnel, and parents (Reckase, 2000; Loomis & Bourque, 2001). These panelists are also balanced on other important demographic characteristics such as ethnicity, age, and geographic region. Additional discussion of the processes, qualifications, and methods for selecting standard setting participants can be found in Raymond and Reid (2001).

This fourth step also involves selecting a standard setting design. Specifically, the standard setting design defines how the panelists are divided and used in the standard setting process. This includes decisions such as whether panelists are to be divided into two or more groups in order to independently replicate the standard setting. As part of the design, decisions also need to be made about whether and how panelists are allowed to interact at the standard setting meeting.

NAEP has used various standard setting designs in different standard settings. Typically, panelists in NAEP standard setting are divided into two separate groups to replicate the standard setting. Within each group, panelists are often organized into four or five smaller groups that work together and talk about how they arrived at their standard setting judgments, the meaning of the PLDs and labels, and item and test content.

Another important consideration at this stage of the process relates to the choice of a facilitator. In most circumstances, the facilitator is an experienced psychometrician with a good understanding of the methods and models that underlie the assessment. In NAEP standard setting, psychometricians from ACT serve as facilitators for the standard setting meeting. These considerations are important because as Fitzpatrick (1989) points

out the social interactions of panelists with each other and the facilitator have the potential to influence the estimated cut scores.

1.1.5 Step 5: Train Panelists to Use the Standard Setting Method

Next, panelists are trained on how to use the standard setting method to make judgments about the location of cut scores. In NAEP, this step involves giving them an overview of the standard setting method and the procedures to follow when providing their standard setting judgments as well as an explanation of the answer keys and scoring rubrics, and a discussion of the test questions. An important part of this step is to help panelists with conceptualizing examinees that just possess the knowledge, skills, and abilities to meet the particular standard represented in the PLD, minimally competent examinees (MCEs). In other words, when providing their judgments, panelists are trained to conceptualize MCEs at each cut score. Coming to this common understanding of MCEs often involves group discussions of PLDs and provides panelists with the opportunity to refine PLDs so that they are more applicable to students with whom they are familiar.

At this stage, panelists are often asked to take the assessment under the same conditions as examinees would take it on the actual day of assessment. In NAEP standard setting, panelists take one to three blocks of items from the NAEP item pool. Panelists, then, are usually given the opportunity to practice the standard setting method on a set of sample items similar to actual items in the operational standard setting. They typically then are asked to discuss their experiences with the facilitator and the other panelists so that panelists feel comfortable with all aspects of the standard setting procedure.

Moreover, this step helps to ensure that panelists can perform the standard setting task as it was intended by the policy board.

1.1.6 Step 6: Collect Panelist Ratings

After receiving the training mentioned above, panelists independently provide their judgments of where they think that cut scores should be set. In NAEP, these standard setting judgments are recorded on rating sheets by the panelists. For example, a panelist might be asked to indicate probability ratings for a set of items for the Angoff method (Angoff, 1971), to indicate page numbers where their bookmarks might be located for the Bookmark method (Lewis et al., 1996), or to indicate which students would be classified into what performance categories for the Contrasting Groups method (Berk, 1976).

1.1.7 Step 7: Compile Panelist Ratings and Obtain Cut Score Estimates

Each panelist's ratings are then used to determine their individual prescribed cut score. Then cut score estimates for all panelists are aggregated, usually using the mean or median of panelists' ratings, and converted to an overall cut score estimate for the group of panelists. In NAEP, the mean and median of the individual panelists have been used to determine an overall cut score estimate for the group of panelists with different standard setting methods.

1.1.8 Step 8: Provide Feedback and Facilitate Discussion

After obtaining an aggregate cut score estimate, the panelists are provided feedback on their individual ratings and those of other panelists. The purpose of this step is to ensure that panelists are comfortable with their judgments and that there were no egregious errors or misunderstandings. The feedback process can vary between rounds and implementations of standard setting procedures. Reckase (2001) provides a good summary of different feedback approaches given to panelists which are:

- 1) Rater location feedback that shows the location of the panelist's cut scores in relationship to the cut scores of the other panelists.
- 2) Consequences data that provides information on the percentage of students that would exceed cut scores at specific locations on the scoring continuum, also known as the percent above cut score (PAC).
- 3) Whole booklet feedback that shows actual samples of student work at different locations along the score scale.
- 4) Internal consistency feedback that gives panelists information about how their standard setting judgments align with a specific level of test performance or scoring model.
- 5) p -value feedback that shows the item difficulty of test items either for the whole population or conditional on their cut score estimates.
- 6) Domain score feedback that shows the expected performance of examinees in specific content domains.

- 7) Other assessment data, such as information that shows how examinees at different grade levels would perform or where the cut scores had been set on similar assessments.

Each of these seven types of feedback have been used in NAEP standard setting in one form or another depending on the standard setting method that was applied.

After reviewing this information, the facilitator often leads the panelists in a discussion about their experiences using the standard setting procedure and whether they think the cut score estimates are reasonable. The feedback and information from the discussion is used in the next round to make new cut score estimates.

1.1.9 Step 9: Conduct Panelist Evaluations

The last part of any particular round of the standard setting process is to conduct an evaluation of the panelists' experiences and to gather opinions about the ratings they provided in that round. Typically, panelist evaluations are in form of a survey that asks specific questions about whether panelists felt they were able to perform the standard setting process effectively and whether the procedures were explained in sufficient detail. In NAEP standard setting, four to seven process questionnaires are included at various points in the standard setting meeting. These questionnaires contain many common elements and typically include both opened-ended and Likert scale questions. An example of questions asked on panelist evaluations as part of a standard setting can be found in Cizek and Bunch (2007). This information serves the essential role of documenting the standard setting process. The information can also be used to refine the standard setting procedure for succeeding rounds.

A round of the standard setting process involves completing steps 6 through 9 (i.e., collecting panelist ratings, compiling ratings and obtaining cut score estimates, providing feedback and facilitating discussion, and conducting panelist evaluations). Most standard setting processes involve several rounds of standard setting before arriving at the final cut score estimates. Hence, the arrow connecting step 9 to step 6 in Figure 1.1 illustrates the repetition of these steps in the process. In NAEP, the standard setting process typically uses three or four rounds of ratings.

1.1.10 Step 10: Prepare Technical Documentation and Validity Evidence

After completion of the standard setting meeting, the individuals who ran the meeting typically write technical reports that document how the standard setting meeting was run, the procedures used in estimating cut scores, and any problems encountered during the process. Any information from special studies that were performed as part of the standard setting meeting is also documented. In addition, detailed statistical analyses of the panelist evaluation surveys are conducted and documented. In NAEP, special studies, technical reports, and process reports are written by ACT and given to NAGB after the standard setting meeting.

The goal of this stage in the process is to collect specific information that can be used when attempting to make validity arguments to support or refute the cut score estimates. This information is used to make a recommendation to the policy board about the reasonableness of cut score estimates and whether or not these estimates should be adopted. Since one of the main purposes of this dissertation is to propose a new framework for evaluating standard setting, a more detailed discussion of technical

documentation and validity evidence that has been collected as part of standard setting processes is provided in Chapter 2.

1.1.11 Step 11: Review Documentation and Determine Final Cut Scores

In the last step, the policy board reviews the technical documentation and validity evidence compiled from the standard setting meeting and determines where the final cut scores should be placed. In NAEP standard setting, NAGB reviews this information and decides whether they want to accept or change the cut estimates provided by the panelists. The decision to adopt or reject the cut scores indicated by the panelists is a policy decision based on information from the standard setting meeting and other important political, economic, and social factors. One possible reason that a policy board might choose to change the cut scores from those suggested by the group of panelists is if they felt that too many examinees would be passing and this would be viewed as making the test too easy. Conversely, if the cut scores would result in an unreasonably low number of examinees passing, the policy board may decide to lower the cut scores. In most cases, aggregated panelist cut scores are adopted and implemented operationally.

1.2 Angoff and Bookmark Methods

As explained previously, an important step in the standard setting process is selecting a standard setting method. The focus in this dissertation is on two test-centered standard setting methods, the Bookmark method (Lewis, et al., 1996) and the Angoff method (Angoff, 1971). These methods are among the most popular methods for operationally setting cut scores and have been used in various forms to set standards on

NAEP as well as in other testing programs. These two methods can be classified as derivatives of the original Angoff method.

1.2.1 Angoff Method

The original Angoff (1971) method is among the most researched and applied standard setting methods (Mehrens, 1995; Brandon, 2004) and was developed from a footnote in Angoff's chapter in *Educational Measurement*. The Angoff method asks panelists to provide a probability judgment that a minimally competent examinee (MCE)—an examinee that just possesses the necessary skills, knowledge, and ability to meet a specific standard— would get each item correct. Each of these probability judgments are summed to arrive at the cut score for an individual panelist and the average across panelists is then used as the cut score for the assessment.

Numerous variations of the original Angoff procedure have been developed in response to the differing needs of testing and assessment programs. Some recent variations of the Angoff method are the Extended Angoff method (Hambleton & Plake, 1995), the Yes/No Method (Angoff, 1971; Impara & Plake, 1997), the Angoff method with Mean Estimation (described in Reckase, 2000), the Item Score String Approach (Bay, 1998; Reckase & Bay, 1998), the Reckase method (Reckase, 1998) and the Direct Consensus method (Sireci, et al., 2004). One might also classify the Basket Procedure (Verhelst & Kaftandjieva, 1999) and the Jaeger Method (1982, 1989) as Angoff derivative methods, although the conceptualization of a MCE differs between these two procedures and some of the other Angoff variations.

1.2.2 Bookmark Method

If one takes a general and liberal view of the Angoff procedure, one might also classify the Bookmark method (Lewis, et al., 1996) and its variants (i.e., Item Map; Shen, 2001), Item Mapping (Wang, 2003), the modified Bookmark method (Buckendahl, et al., 2002), the Mapmark method (Schulz & Mitzel, 2005; in press), and the Single Passage Bookmark (Skaggs, et al., 2007)) as Angoff derivatives. Specifically, Bookmark procedures can be viewed as Angoff derivatives since panelists have to conceptualize the probability of a MCE obtaining a score point that is greater than or equal to the response probability (RP) criterion (e.g. a probability level) when providing standard setting judgments. The RP criterion serves two essential roles in the Bookmark procedure. The first is to determine the θ location of the items in the ordered item booklet (OIB). In this case, the θ location where each item is equal to probability level specified by the RP criterion is located and the items are ordered based these θ locations. The second use of the RP criterion is as the probability threshold that panelists conceptualize as they move through the OIB. That is, each panelist who performs the Bookmark procedure moves through the OIB asking themselves the question of whether or not the MCE should obtain that score or higher with probability greater than the specific probability level. Notice how panelists are still required to provide a probability judgment for each item, but the probability judgment is in relation to the threshold specified by the RP criterion.

The conceptualization of the Bookmark method as an Angoff method hybrid is not widely held and some scholars would argue that the structure and cognitive task asked of panelists are completely different between the two (Lewis, et al., 1996; Schulz, 2006). Scholars who ascribe to this view believe that the ordering of the items into an

OIB, as they are in Bookmark procedures, changes the task from providing probability ratings to locating a place along a difficulty scale when selecting a cut score (Schulz, 2006). The direct relationship between Angoff and Bookmark methods is explained in greater detail below.

1.2.3 Overview of Angoff Derivative Methods

An overview of most of the commonly used Angoff derivative methods including the stimulus, types of test items, conceptualization of a MCE, the rating method, and the methods for deriving the cut scores can be found in Table 1.2. Oftentimes, “Angoff” method is used as a label for many of the standard setting processes in Table 1.2. For example, some researchers refer to the Yes/No method directly as an “Angoff” method (Davis, et al., 2008). However, the original Angoff method and the Yes/No method are different. A closer examination of any standard setting method in Table 1.2 allows one to see how these methods compare to the original Angoff procedure.

Table 1.2: Angoff Derivative Standard Setting Methods

<i>Method</i>	<i>Stimulus</i>	<i>Types of Test Items</i>	<i>Conceptualization of Minimally Competent Examinee (MCE)</i>	<i>Rating Method</i>	<i>Method for Deriving a Panelist's Cut score</i>
Angoff (1971)	Individual Test Items	Dichotomous Items	An examinee that just possesses the necessary skills, knowledge, and ability to meet the standard	Probability that the MCE will answer the item correctly. Another variation is to conceptualize 100 MCEs and indicate how many of the examinees should answer the item correctly.	Sum of the probability ratings across test items Map through the test characteristic curve to obtain cut score in θ metric
Angoff with Mean Estimation (described in Rackase 2000 & Loomis & Bourque, 2001)	Individual Test Items	Dichotomous and Polytomous Items	Same as Angoff	The panelists provide an estimate of the mean score a MCE should be expected to obtain on each of the items. Panelists might also indicate the average score that 100 MCEs should have on the item.	Sum of item ratings across test items Map through the test characteristic curve to obtain cut score in θ metric
Yes/No Method (Angoff, 1971; Impara & Plake, 1997)	Individual Test Items	Dichotomous	Same as Angoff	Panelists provide a 1 or 0 to indicate whether or not they believe the MCE should or should not get the item correct. Sometimes the instructions for the method ask panelists to indicate a 1 if they believe a MCE has a greater than 50 percent chance of getting the item correct and 0 otherwise.	Sum of the ratings for the individual test items Map through the test characteristic curve to obtain cut score in θ metric

Table 1.2 (cont'd)

<i>Method</i>	<i>Stimulus</i>	<i>Types of Test Items</i>	<i>Conceptualization of Minimally Competent Examinee (MCE)</i>	<i>Rating Method</i>	<i>Method for Deriving a Panelist's Cut score</i>
Extended Angoff (Hambleton & Plake, 1995)	Individual Test Items	Polytomous Items	Same as Angoff	Panelists indicate the score on the item that they think a MCE is most likely to receive.	Sum of the ratings across test items Map through the test characteristic curve to obtain cut score in θ metric
Item Score String Estimation (Bay, 1998; Reckase & Bay, 1998)	Individual Test Items	Dichotomous and Polytomous Items	Same as Angoff	Panelists indicate the score on the item that they think a MCE is most likely to receive.	The maximum likelihood estimate (or the EAP or MAP estimate) of the item score string is taken as the cut score.
Direct Consensus Method (Sireci et al., 2004)	Clusters of items organized around a common strand or content area	Dichotomous and Polytomous Items	Same as Angoff	The panelists indicate the number of items in each cluster that they believe a MCE should be able to answer correctly. In the last round, panelists are asked to come to a consensus as to the number of items a MCE should answer correctly in each cluster.	The sum of the number of items that would be answered correctly across all of the content clusters. Map through the test characteristic curve to obtain cut score in θ metric.

Table 1.2 (cont'd)

<i>Method</i>	<i>Stimulus</i>	<i>Types of Test Items</i>	<i>Conceptualization of Minimally Competent Examinee (MCE)</i>	<i>Rating Method</i>	<i>Method for Deriving a Panelist's Cut score</i>
Reckase Method (Reckase, 1998)	Individual Test Items and a Reckase Chart in round 2 and later.	Dichotomous and Polytomous Items	Same as Angoff	Same as Angoff in the first round. In second and later rounds, panelists indicate their probability ratings with the assistance of the Reckase chart or draw a line to indicate where they would like to place their cut score.	Same as Angoff in the first round. In second and later rounds, the cut score is either the sum of probability ratings or the value of test characteristic curves at the θ value indicated on the Reckase chart Map through the test characteristic curve to get the cut score in the θ metric.
The Basket Procedure (Verhelst & Kafandjieva, 1999)	Individual Test Items	Dichotomous Items	An examinee at a specific performance level	Panelists indicate whether a test taker at a specific performance level should or should not be able to answer the test item correctly.	The total number of items judged to be correct at next lowest performance level. Map through the test characteristic curve to get the cut score in the θ metric.

Table 1.2 (cont'd)

<i>Method</i>	<i>Stimulus</i>	<i>Types of Test Items</i>	<i>Conceptualization of Minimally Competent Examinee (MCE)</i>	<i>Rating Method</i>	<i>Method for Deriving a Panelist's Cut score</i>
Jaeger Method (Jaeger, 1982; 1989)	Individual Test Items	Dichotomous Items	Every examinee taking the assessment	Panelists indicate whether every examinee should or should not be able to answer the test item correctly using a yes or no. The method is only applied to set a single cut score.	The total number of items that every examinee should answer correctly.
Bookmark (Lewis et al., 1996)	A booklet of ordered test items based on a response probability criterion.	Dichotomous and Polytomous Items	Same as Angoff	Panelists move through the ordered item booklet keeping in mind the response probability used to order the items. Panelists place a bookmark in between the last item that the MCE should be able to answer correctly with the specified response probability and the first item that they should not be able to answer correct at the specified response probability.	The θ value for the item that preceded the bookmark. In the true score metric, it is the value of test characteristic curve at the θ value for the item that precedes the bookmark.
Modified Bookmark (described in Buckendahl, et al. 2002)	A booklet of items ordered on a response probability criterion. Sometimes the items are ordered by the p values of the items	Dichotomous and Polytomous Items	Same as Angoff	Same as Bookmark	The sum of the number of test items before the bookmark is taken as the cut score. Map through the test characteristic curve to obtain cut score in θ metric.

Table 1.2 (cont'd)

<i>Method</i>	<i>Stimulus</i>	<i>Types of Test Items</i>	<i>Conceptualization of Minimally Competent Examinee (MCE)</i>	<i>Rating Method</i>	<i>Method for Deriving a Panelist's Cut score</i>
Mapmark (Schultz & Mitzel, 2005, in press)	A booklet of ordered test items, a map that organizes the items into the content domains, and a map showing the location of the items along the score scale. Sometimes panelists also receive whole booklet feedback.	Dichotomous and Polytomous Items	Same as Angoff	Same as Bookmark for the first round. In the second and later rounds, panelists mark their ratings on an item map or in the ordered item booklet. Panelists also receive information on the expected performance in each of the domain, domain score feedback (Schulz, et al., 1999), and can use this feedback to help make their standard setting judgments. Another variation of the procedure is to provide panelists with samples of student performance at each of the score scale points instead providing the expected performance in the domains.	Same as Bookmark for the first round. In subsequent rounds the cut scores corresponds to the θ value on the chart or the item that precedes the bookmark if the panelist indicates their cut score in the ordered item booklet.
Item Map (Shen, 2001)	Several maps of items that are ordered by the item difficulties and separated according to different content dimensions.	Dichotomous Items that are calibrated with the Rasch model.	Same as Angoff	Panelists draw a line on each item map to separate the items that a student should have mastered from the items that they should not have mastered in content dimension. Mastery is defined as having a greater than 50 percent chance of answering the question correct.	The cut score for each item map is the item difficulty for the item directly below the line drawn by the panelist. The average of the item difficulties across the item maps is the estimated cut score.

Table 1.2 (cont'd)

<i>Method</i>	<i>Stimulus</i>	<i>Types of Test Items</i>	<i>Conceptualization of Minimally Competent Examinee (MCE)</i>	<i>Rating Method</i>	<i>Method for Deriving a Panelist's Cut score</i>
Item Mapping (Wang, 2003)	A histogram chart of all of the items based on the estimated difficulty of the items.	Dichotomous items that are calibrated with the Rasch model.	Same as Angoff	Panelists look at the item map and ask themselves whether the MCE has a greater than fifty percent chance of answering the item correctly. If the answer is yes they move to the next item, if the answer is no that item becomes the preliminary cut score. Panelists are then given a graphic showing the location of each of the items and the probability of answering the items correctly. If the graphic does not match what a MCE should be able to do, then panelists are instructed to select a new cut score. A new graphic is given to them and they asked whether this cut score is consistent with what a MCE should be able to do.	The θ estimate of the location of the cut score in the last graphic that they select.
Single Passage Bookmark (Skaggs, et al., 2007)	A booklet of ordered test items for each separate reading passage.	Dichotomous and Polytomous items related to a reading passage.	Same as Angoff	Panelists move through the ordered item booklets keeping in mind the response probability used to order the items. Panelists place a bookmark in between the last item that the MCE should be able to answer correctly and the first item that they should not be able to answer correctly at the specified response probability for each reading passage.	The median of θ values for the item that preceded the bookmark across the full set of passages.

Each of the standard setting methods described in Table 1.2 consists of asking panelists to provide judgments of how an examinee at a specific performance level should hypothetically perform on test items. The main differences in each of the methods can be described in terms of how panelists are asked to provide ratings to the test items, whether panelists are asked to conceptualize a specific RP criterion or not, whether the items are presented individually or as a set, whether the items are ordered (i.e., from easiest item to hardest item) or not, the types of items that are rated (i.e., dichotomous, polytomous, or both), and how the cut scores are determined from the ratings of the panelists.

For example, the Bookmark method differs from the Angoff method in that it orders items from easiest to hardest based on a particular RP criterion into a set of items called an OIB while for the original Angoff method the items are not ordered using a RP criterion. Instead, panelists are asked to rate the items in the order that they would appear on the assessment. In most applications of the Bookmark procedure, a RP criterion of 0.50 or 0.67 is used (Hyunh, 2006). The decision to use a particular RP criterion is a policy decision that is made by the policy board that sets the standards. In practice, RP values ranging from 0.5 to 0.8 have been applied with the Bookmark method (Zwick, et al., 2001)

The RP criterion defines the specific probability level of obtaining a score at that particular level or higher for a MCE and it is used to order the items in the OIB. For example, if the RP criterion is 0.67 and all the items are dichotomous, then for each item the specific value on the score scale (the θ -scale in IRT) associated with getting the item correct 67 percent of the time would be determined.

The items would then be placed into a booklet based on the score values that correspond to getting the item correct 67 percent of the time from lowest score scale value (the easiest item) to the highest score scale value (the hardest item). Panelists would then be asked to proceed through the booklet of items asking themselves the question of whether an examinee just above the standard should or should not be able to answer the item correctly 67 percent of the time. A panelist places a bookmark in between the last item that they believe a student who is just above the standard should be able to answer correctly and the first item the student should not be able to answer correctly at the 67 percent level. The cut score is determined by finding the score scale value that corresponds to getting the item preceding the bookmark correct 67 percent of the time. The panelist repeats this process for each cut score that they need to set.

Notice how, in theory, this is an Angoff procedure where the items are ordered according to the RP criterion and the panelists make a probability judgment of whether the probability of answering the item correctly for the MCE exceeds a threshold specified by the RP criterion. The panelists do not actually have to indicate the probability, but in theory they are supposed to assess whether the probability of getting the item correct exceeds the threshold.

The focus in this dissertation is on first round of the Mapmark method where the method is essentially the regular Bookmark procedure and the Angoff method with Mean estimation. Each of these methods is outlined in Table 1.2. These methods are explained in greater detail in later sections of this dissertation.

1.3 Formulation of Problem

An inherent assumption made in the application of many standard setting procedures is that the panelists understand the task they are asked to perform and are able to carry out the procedure accurately. In essence, the whole system of educational accountability is critically dependent on the reasonableness of cut scores and their efficacy in representing achievement goals reflective of educators' and policy makers' conceptualizations of the knowledge, skills and abilities required by students in order for them to be successful academically. Unfortunately, panelists do not always carry out standard setting correctly (Cizek, 2001; Kane, 2001). In addition, the ways that some of the methods are implemented can introduce problems in determining the location of the cut score. These potential problems have been recognized by some researchers and have led to pointed criticism and debate about the mechanisms and methods for determining cut scores (Cizek, 2001).

One prominent example of a standard setting critique was of the initial methods used to set cut scores on NAEP (Shepard et al., 1993; Shepard, 1994; Linn & Shepard, 1997). Critics argued that the methods were overly complex for panelists to use and that in practical settings panelists struggled to understand the rating process and to provide consistent ratings. Specifically, Shepard et al. (1993) and Shepard (1994) suggested that panelists who used the Angoff procedure often rated different item types inconsistently. They observed that panelists often viewed polytomous items to be more challenging than dichotomous items and that if cut score were established using different item types there tended to be large disparities in rater judgments. They also discovered that the probability ratings given by panelists to individual items were not perfectly related to the difficulties

of the test items. Their research suggested that there could be some large potential biases and inaccuracies in cut scores when applying the Angoff procedure, although they did not specifically quantify the potential biases.

Over the last decade, there has been additional criticism of the Angoff procedure. Schulz (2006) agreed with Shepard et al. (1993) and Shepard (1994) that the Angoff procedure suffers from regression to the mean of the probability scale. He also indicated that panelists tended to round their ratings. Impara and Plake (1997, 1998) along with Plake and Impara (2001) also argued that panelists are not good at estimating probability in general and suggested an alternative standard setting method, the Yes/No method, that they believe simplifies the process considerably.

Criticism of other standard setting methods has included the Bookmark procedure (Berk, 1996), which is currently the most widely used standard setting method in state testing programs (CCSSO, 2001; Karantonis & Sireci, 2006). Specifically, researchers have pointed out that the Bookmark method suffers from RP indeterminacy (Haertel & Lorie, 2004). That is, there is not one unique RP criterion that underlies test performance. In particular, it is possible to design the OIB used in the Bookmark method based on any RP level between zero and one hundred. Two important observations in this context is that there is the possibility for the order of the items in the OIB to change and for the cut scores to be different if the RP criterion is modified (Kolstad, et al., 2001; Skaggs & Tessema, 2001; Kolstad, 2002; Beretvas, 2004; Williams & Schulz, 2005). Cizek and Bunch (2007) have also observed that in order for the Bookmark procedure to be accurate there should be a large number of items in the OIB that are near the location where the panelist intends to set their cut score.

Much of the debate and criticism of the standard setting procedures (Shephard, et al., 1993, Shephard, 1994; Hambleton, et al., 2000) used in NAEP and other settings are disagreements about whether the cut scores are accurate representations of the intended cut scores of the panelists and/or whether the evidence collected to evaluate the quality of the standard setting actually provides indications that the standard setting procedures did or did not work effectively. Specifically, researchers have disagreed about what criteria and guidelines should be used to evaluate standard setting (Cizek, 1996; Kane, 1994, 2001). An important observation made in the literature is that most of the evidence collected to date can rule out a standard setting method, but it can never rule it in (Hambleton & Pitoniak, 2006). This observation stems from the fact that until recently no framework had been proposed that actually allowed researchers and practitioners to investigate the ability of standard setting methods to produce the hypothetical cut scores that a panelist wanted to set. Consequently, until the proposal of the psychometric framework suggested by Reckase (2006a) very few systematic investigations of standard setting processes and methods have been performed (Engelhard, in press), and the ones that have been performed did not provide a clear indication of whether or not the standard setting process was effective.

1.4 Purpose

Therefore, the purpose of this dissertation is to further develop a framework for standard setting that can be used to evaluate any IRT based standard process in operational or simulated settings for potential biases in cut scores judgments. In addition, an application of how the new framework can be used to evaluate standard setting

procedures is illustrated using NAEP data. In this dissertation, I will show how the newly-developed framework, which extends Reckase's psychometric theory (Reckase, 2006a) aided by construct maps (Wilson, 2005), can be used to create indices to evaluate the Angoff and Bookmark procedures for potential biases since these methods are applied most often operationally. Therefore, this study will seek to address the following research questions:

- 1) How can the new comprehensive framework based on extending Reckase's psychometric theory be used to evaluate and improve standard setting?
- 2) What are the potential cut score biases produced from using the Bookmark or Angoff methods in NAEP standard setting and how comparable are biases between the two methods?

In Chapter 2, previous approaches to evaluating the reasonableness of cut scores from standard setting are reviewed including approaches based on the multifaceted Rasch model (MRM) (Engelhard, in press), making a validity argument for or against the cut scores (Kane, 2001), and Reckase's (2006a) psychometric theory. In Chapter 3, a new comprehensive theoretical framework for evaluating standard setting methods is presented that extends Reckase's (2006a) psychometric theory in conjunction with construct maps (Wilson, 2005). Chapter 4 demonstrates how the new theoretical framework for evaluating standard setting can be used to develop models and indices for evaluating the Angoff and Bookmark methods for potential biases and inconsistencies. Applications of the new statistical models and indices to operational standard setting data from NAEP are presented in Chapter 5. The implications of the results from the investigations of NAEP standard setting are also discussed in Chapter 5. Finally, in

Chapter 6 the significance of the new theoretical framework for future standard setting practice is discussed and some areas for future research are presented.

CHAPTER 2

LITERATURE REVIEW

An overview of the standard setting process and the standard setting methods that are the focus of this dissertation was provided in Chapter 1. This chapter's goal is to review previous efforts and proposed frameworks for evaluating the quality of standard setting. In particular, Kane's (1994, 2001) validity framework for evaluating the reasonableness of cut scores, Engelhard's (Engelhard & Anderson, 1998; Engelhard, in press) Rasch based framework for evaluating standard setting judgments, and Reckase's (Reckase, 2006a) psychometric theory for evaluating standard setting procedures are explained and reviewed. These frameworks are reviewed because they lay the groundwork for the development of the new framework developed in this dissertation. The relationship of this study's newly formulated framework to prior research is explained. Finally, previous empirical comparisons of the Angoff and Bookmark standard setting methods are reviewed.

2.1 Kane's Validity Framework for Standard Setting

Kane's (1992, 1994, 2001, 2006) validity framework for standard setting is by far the most common framework for evaluating the quality of standard setting. He argues that one of the most essential components of any psychometric or measurement endeavor is the evaluation of how its results are used and interpreted. Kane, following the work of Messick (1988, 1989), believes that measurement validity is critical and the goal of the researcher or practitioner is to build an argument, in much the same way as arguments are

built in court cases, for or against the intended uses and interpretations of the test results. In the standard setting context, Kane's (1994, 2001) framework for evaluating panelist judgments is to build an argument for or against the use of cut scores in much the same way as validity arguments are made in other areas of measurement. In his chapter in *Setting performance standards: Concepts, methods, and perspectives*, Kane (2001) suggests three types of evidence that one can collect when attempting to make a validity argument in support of the use of cut scores. These three types of evidence include: collecting procedural validity evidence, collecting internal validity evidence, and collecting external validity evidence from the standard setting. These same three categories of validity evidence are described in a dissertation by Pitoniak (2003) and a review chapter by Hambleton and Pitoniak (2006). The review of these three types of validity evidence and how they can be used to evaluate standard setting follows from the work of these authors.

2.1.1 Procedural Evidence

Procedural validity evidence consists of collecting information about the procedures used in establishing the standards and the corresponding cut scores. Examples of procedural evidence include:

- 1) Explicitness - collecting information about the degree to which the standard setting was clearly defined prior to implementation (van der Linden, 1995),
- 2) Practicability - collecting information about how easy it was to conduct the standard setting procedure and how much the procedure makes sense to the general public (Berk, 1986),

- 3) Implementation of Procedures - collecting information about the extent to which the procedures were systematic and thorough (Kane, 1994; 2001),
- 4) Panelist Feedback - collecting information about the extent to which panelists felt comfortable with the process and the result of the standard setting (Kane, 1994; 2001), and
- 5) Documentation - collecting evidence of how well the standard setting methods and procedures are documented for evaluation purposes (Cizek, 1996; Hambleton, 1998, Mehrens, 1995).

Collecting this information is important when examining a standard setting process since it would be difficult to justify the cut scores produced if the procedures used to derive them were unsystematic, poorly documented, and hard to understand. However, it is easy to see that these types of evidence by themselves do not establish the correct functioning of the standard setting process or the reasonableness of cut scores. For example, panelists may feel comfortable with the standard setting procedure, but they could be performing it in a manner that results in unreasonable cut scores. The standard setting procedure could also be well documented, explicit, and practical, but not produce reasonable cut scores. Clearly, even though this information is important to collect as part of standard setting process, it is not sufficient for evaluating the quality of cut score estimates.

2.1.2 Internal Evidence

Internal validity evidence of standard setting quality is established by collecting evidence to support or refute the consistency of cut scores and panelist ratings during standard setting. Examples of internal evidence include:

- 1) Consistency within Method - collecting information about how well the cut score estimates would compare to each other if the standard setting was replicated (Cizek, 1996; Kane, 1994, 2001),
- 2) Intrapanelist Consistency - collecting evidence of each panelist's ability to consistently rate item difficulties across standard setting rounds (Berk, 1996; Cizek, 1996, van der Linden, 1982),
- 3) Interpanelist Consistency - collecting evidence of item rating and cut scores consistency across panelists (Berk, 1996; Cizek, 1996; Jaeger, 1989),
- 4) Standard Error of the Cut Score - examining cut score precision (Kane, 2001), and
- 5) Other Measures - examining cut score consistency across item types, content areas, and/or cognitive processes (Kane, 1995, Shepard, et al., 1993).

These types of internal validity evidence provide an indication of panelists' ability to provide systematic ratings in standard setting. These systematic ratings are desirable because they can indicate whether a panelist or group of panelists have provided erratic and inconsistent judgments, which may impact the meaningfulness of the cut score that is estimated. However, one issue with these types of evidence is that the consistency or precision of a panelist or group of panelists is not conceptualized in terms of the cut scores that the panelist or group of panelists had in mind when they provided their

standard setting judgments. Without this link, there is the possibility for panelists to be precise and consistent, but the precisely estimated cut score may be different than the cut score that a panelist had in mind when providing their standard setting judgments. For example, it is possible for panelists to estimate similar cut scores (which would be viewed as high quality interval validity evidence), but the cut scores could be biased in a similar fashion from panelists making the same type of errors in the standard settings. Again, this information is highly informative but a clear link between this evidence and the quality of the cut scores is often nonexistent in the evaluation.

2.1.3 External Evidence

External validity evidence of standard setting quality consists of examining the relationship of the cut scores to other important external criteria (e.g., student grades, performance on other assessments, other research studies). Examples of external evidence include:

- 1) Comparisons to Other Standard Setting Methods - collecting evidence of the similarity of cut scores when applying different standard setting methods (Jaeger, 1989; Kane, 1994, 2001),
- 2) Comparisons to Other Sources of Information - examining the relationship of decisions made based on the cut scores to grades or performance on other tests (Berk, 1996; Kane, 1994, 2001; Shephard, et al., 1993), and,
- 3) Reasonableness of Performance Levels - examining the extent to which the passing rate and the cut score appears to be plausible for the examinee population (Kane, 1998, 2001).

Again, these sources of evidence are important, but not sufficient for ensuring that the standard setting procedure is functioning appropriately. For example, two different standard setting methods could produce highly similar results, but both procedures could be biased in the same direction. Further, one might expect differences between the decisions based on cut scores and decisions based on other external criteria since these external criteria could be measuring different aspects of student ability than those represented by the cut scores. Concerns could also be raised about using the passing rate to examine the quality of standard setting since this would appear to defeat one of the main purposes of setting standards on criterion-referenced tests. Namely, one of standard setting's main purposes is that assessment standards represent what stakeholders think examinees should know and be able to do instead of arbitrarily passing a specific proportion of students from an examinee population. In other words, using the passing rate independent of assessment standards does not adequately evaluate standard setting quality.

2.1.4 Strengths and Weaknesses of Kane's Standard Setting Validity Framework

Kane's approach has some strengths and weaknesses for evaluating standard setting. One of the strengths of this approach is that the combination of standard setting components allows the evaluation approach to be presented in such a way that standard setting process advantages and disadvantages can be weighed against each other. Additionally, the different types of evidence collected with this approach can provide some indications of potential problems in standard setting. For example, if the panelists were unable to understand the standard setting process or if the same standard setting

procedure resulted in drastically disparate results with two equivalent groups of panelists, this would provide evidence that the standard setting process might not be working effectively.

However, this framework is not sufficient for evaluating standard setting procedures because it does not address the procedure's robustness in recovering a panelists' intended cut scores. That is, there is no indication of the quality of panelist standard setting judgments in relationship to the cut scores that panelists had in mind when they provided their judgments. The quality of panelists' judgments in relationship to an intended cut score is a fundamental factor which cannot be directly addressed in the current validity argument approach to evaluating standard setting since the framework does not assume there is an intended cut score that a panelist intends to set. Therefore, there are no statistical procedures available for quantifying the potential biases or inconsistencies that may be present in intended cut score estimates under this framework. As Hambleton and Pitoniak (2006) correctly surmise, this approach to evaluating standard setting can only provide an indication that a standard setting method may not be working; it can never indicate whether standard setting was actually precise, accurate, or effective.

2.2 Engelhard's Rasch Evaluation Framework

The second commonly used approach for evaluating standard setting is Engelhard's framework (Engelhard & Anderson, 1998; Engelhard & Stone, 1998; Engelhard, 2007, in press, Caines & Engelhard, 2009). This framework applies the

multifaceted Rasch model (MRM) to standard setting judgments arising from various standard setting procedures. The MRM is represented as:

$$\text{Ln}(P_{nijk} / P_{nijk-1}) = \theta_n - \delta_i - \omega_j - \tau_k, \quad (2.1)$$

where P_{nijk} = probability of panelist n giving rating k on item i for cut score j ,
 P_{nijk-1} = probability of panelist n giving rating $k-1$ on item i for cut score j ,
 θ_n = judgment of MCE required to pass for panelist n ,
 δ_i = judgment of difficulty for item i ,
 ω_j = judgment of cut score for round j ,
 τ_k = judged threshold of rating category k relative to rating category $k-1$.

The MRM in Equation 2.1 models the probabilities of providing ratings in different successive rating categories, the log odds of being in the higher category compared to being in the next lower category, as a function of various facets that might influence the panelist's tendency to provide a rating in the different rating categories.

This framework has several notable advantages when evaluating standard setting: (a) it uses a hypothetical cut score that a MCE is required to pass; (b) it uses various factors that might impact the standard setting judgments (e.g., panelist table groups), including the previous round of standard setting; and (c) it uses psychometric models and indices for evaluation. In this framework, the person performing the evaluation examines estimates of the various factors and the INFIT and OUTFIT statistics from MRM in order to assess the quality of the standard setting.

Engelhard (in press) gives a detailed explanation of how these statistics and estimates can be used to identify various potential issues that might be present in the standard setting process. In particular, he shows how the MRM is useful for identifying:

- 1) Rater Severity - the panelist's tendency to provide higher or lower ratings than they should,
- 2) Halo Effect - the inability of panelists to distinguish between independent and distinct aspects of examinee performance,
- 3) Response Sets (Central Tendency) - the tendency of panelists to over use certain rating categories when they should not, (i.e. panelists over use the middle categories of the rating scale),
- 4) Restriction of Range - the inability of panelists to accurately discriminate among the different performance levels when setting multiple cut scores so that the cut scores are too close to each other,
- 5) Interaction Effects - the different facets (e.g. rounds and table groups) are not independent and additive for an individual or group of panelists, and,
- 6) Differential Facet Functioning - the measurement of the model's different facets are impacted by construct irrelevant variance such that the model is not invariant (e.g. raters from different demographic subgroups of the population are not exchangeable).

Often these issues represent some of the biggest concerns in estimating cut scores. Since the MRM can identify these potential problems using familiar statistics and since the model could be used during the standard setting procedure to provide panelists with

information about their judgments, this gives it significant traction as an evaluation strategy.

However, this technique has some limitations. One limitation is that it uses a Rasch scaling procedure, which may not be the scaling method used on the assessment. This would mean that if the test was scaled with the three-parameter logistic (3PL) model then the scale of standard setting evaluation and the scale of the assessment would be different. This suggests that the hypothetical cut score in the MRM, θ_n , is typically not the same as the one that the panelist intended to set on the assessment, which may create an issue since the concern in the evaluation is often with potential biases and inconsistencies in the hypothetical cut score in the metric of the assessment.

Another issue with Engelhard's framework is that it often requires some recoding and collapsing of panelist ratings (Engelhard & Anderson, 1998). For example, one cannot directly apply the MRM to the probability ratings provided by panelists when using the Angoff procedure since several rating categories would not have any ratings in them. These missing rating categories cause the MRM to have convergence problems during parameter estimation. Unfortunately, the collapsing of rating categories to ensure model convergence often results in the loss of potentially valuable information.

Furthermore, a linking procedure is required if one wants to compare standard setting results at different points in time since the evaluation metric at different time points would not necessarily be the same. Lastly, this framework requires slightly different methods for different standard setting procedures, which would also require the use of sophisticated linking methods to compare different standard setting procedures. For example, the coding schemes and evaluation techniques are somewhat different for

the Bookmark and Angoff methods (see Engelhard, in press; Engelhard & Anderson, 1998). Specific to this study, no procedures exist for linking the evaluation scales of the Angoff and Bookmark methods.

2.3 Reckase's Psychometric Theory

The third and final approach is Reckase's (2006a) psychometric theory for evaluating standard setting judgments. It is similar to Engelhard's approach in the use of an IRT model and the assumption of a hypothetical intended cut score that a panelist would like to set on the assessment. Reckase's approach differs from Engelhard's in that it does not necessarily have to use a Rasch based framework and it allows one to perform the evaluation of standard setting in the metric used to scale the assessment.

Therefore, in order to provide a better understanding of Reckase's framework, the most common dichotomous and polytomous IRT models are reviewed. Then, the relationship of these models to the Reckase's psychometric theory is explained in greater detail. Finally, prior related research is discussed, along with how it provides impetus for the current study.

2.3.1 *IRT Models*

An important basis for Reckase's psychometric theory is unidimensional IRT. IRT is a psychometric framework used for modeling the propensity of obtaining a particular score on a test item as a function of ability and the characteristics of the test item. All parametric IRT models typically have four common assumptions (Hambleton, et al., 1991). These assumptions are:

- 1) Monotonicity (i.e., with increasing ability the probability of obtaining a correct response can never decrease),
- 2) Statistical independence (i.e., once the correct number of abilities have been controlled for, the probability of jointly responding to a set of items is equal to the product of the probabilities of responding to each item individually across all the individuals taking the test),
- 3) Functional form (i.e., the IRT model describes the underlying data), and
- 4) Population and parameter invariance (i.e., the IRT models and parameters for items do not change across populations).

The most popular IRT models are the dichotomous IRT models, the Rasch model (Rasch, 1960), the two-parameter logistic (2PL) model, and three-parameter logistic (3PL) model (Birnbaum, 1968, Lord, 1980). The 3PL model is represented as follows:

$$P_i(\theta) = P(X_i = 1 | \theta) = c_i + (1 - c_i) \frac{\exp(1.7a_i(\theta - b_i))}{1 + \exp(1.7a_i(\theta - b_i))}. \quad (2.2)$$

where θ is a latent unobserved ability, a_i is the slope or discrimination parameter for item i , b_i is the location or difficulty parameter for item i , and c_i is the chance or pseudo-guessing parameter for item i . The 2PL model is a special case of the 3 PL model when the pseudo-guessing parameter is equal to 0 and is given by

$$P_i(\theta) = P(X_i = 1 | \theta) = \frac{\exp(1.7a_i(\theta - b_i))}{1 + \exp(1.7a_i(\theta - b_i))}. \quad (2.3)$$

The Rasch model is written as

$$P_i(\theta) = P(X_i = 1 | \theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}. \quad (2.4)$$

When the test items are not scored as right or wrong, polytomous models are used to model the propensity of obtaining a particular response on the test item as a function of examinee ability and item characteristics. The two most common polytomous models are the partial credit model (PCM; Masters, 1982) and the generalized partial credit model (GPCM; Muraki, 1992). The PCM is represented as:

$$P_{ix}(\theta) = P(X_i = x | \theta) = \frac{\exp \sum_{k=0}^x (\theta - b_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^h (\theta - b_{ik})}, \quad x = 0, 1, \dots, m_i \quad (2.5)$$

where $P_{ix}(\theta)$ denotes the probability of person with ability θ receiving a score of x on the item i , m_i represents the highest possible score for item i , b_{ik} is threshold parameter between category k and category $k+1$, and there are $m_i + 1$ available score categories for the item.

The GPCM is denoted as:

$$P_{ix}(\theta) = P(X_i = x | \theta) = \frac{\exp \sum_{k=0}^x a_i (\theta - b_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^h a_i (\theta - b_{ik})}, \quad x = 0, 1, \dots, m_i, \quad (2.6)$$

where the only difference between Equation 2.6 and Equation 2.5 is the inclusion of a discrimination parameter. Both the PCM and GPCM simplify to dichotomous IRT models when there are only two score categories (Masters, 1982 and Muraki, 1992). Oftentimes, the parameter b_{ik} in the GPCM is represented and estimated as:

$$b_{ik} = b_i + d_k, \quad (2.7)$$

where b_i is an item-location parameter and d_k is a category parameter with the constraints that

$$\theta - b_{i0} \equiv 0, \quad (2.8)$$

and

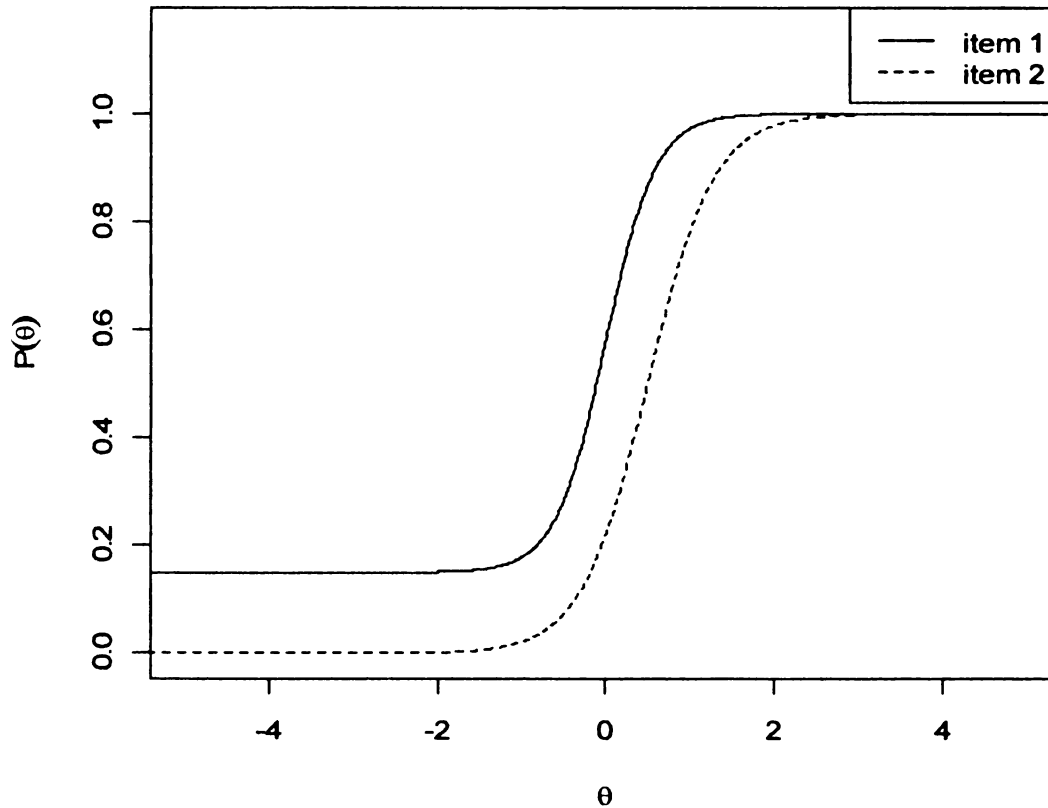
$$\sum_{h=1}^{m_i} d_k = 0. \quad (2.9)$$

The notation in Equations 2.7 through 2.9 is commonly used in the software package PARSCALE (Muraki & Bock, 1991).

There are many other polytomous models besides the PCM and the GPCM such as the graded response model (Samejima, 1969), the nominal response model (Bock, 1972), and the rating scale model (Anderson, 1977; Andrich, 1978). The interested reader is referred to van der Linden and Hambleton (1997) for a discussion of these and other IRT models.

Associated with each of the above IRT models are the concepts of item characteristic curves (ICCs) and test characteristic curves (TCCs). The ICC depicts the expected score on an item as a function of ability. ICCs are useful for comparing item performance of different items. For dichotomous IRT models, the ICC is exactly the same as the item response function for that item. These are Equations 2.2, 2.3, and 2.4 for the 3 PL, 2PL, and Rasch models, respectively. Examples of ICCs for two different 3PL items are shown in Figure 2.1.

Figure 2.1: Example of an Item Characteristic Curve for Two Items



The item parameters for item 1 are $a = 2.0$, $b = 0.0$, and $c = 0.15$ and for item 2 the item parameters are $a = 1.5$, $b = 0.5$, and $c = 0.0$

For polytomous items with more than two score categories, the ICC and the item response functions are not the same. For these items, the ICC is defined as:

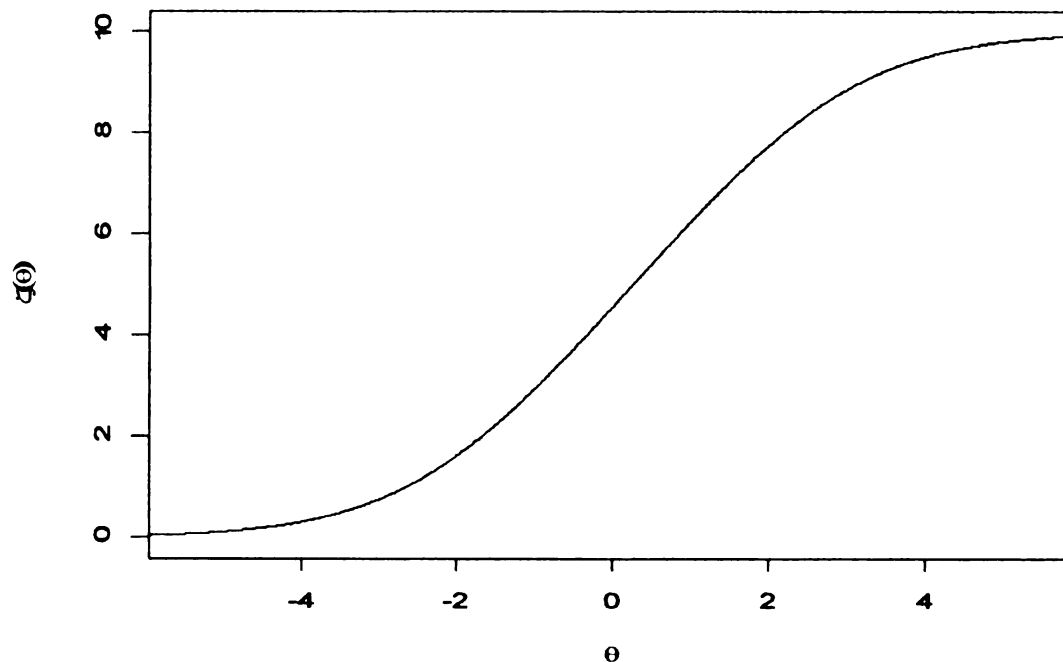
$$E(X|\theta) = \sum_{x=0}^{m_j} x \cdot P_{ix}(\theta), \quad (2.10)$$

The TCC is defined simply as the sum of the ICCs over items. This can be represented as:

$$\xi(\theta) = \sum_{i=1}^n E(X_i|\theta). \quad (2.11)$$

Equation 2.11 relates the overall expected performance on a set of items as a function of ability. The IRT TCCs are useful for comparing the expected performance for different sets of items within and across assessments. The value of the TCC at the particular value of θ is the number-correct true score for that θ value. An example of a TCC for a 10 item test composed of 10 Rasch items whose item difficulties are in increments of 0.5 and range from -2 to 2.5 is shown in Figure 2.2.

Figure 2.2: Example of a Test Characteristic Curve for 10 Rasch Items



2.3.2 Link Between IRT and Reckase's Psychometric Theory

Reckase's (2006a) psychometric theory for standard setting relates the intended cut score that a panelist had in mind when they provided their cut score judgments and the standard setting procedure through IRT models (i.e., the Rasch model, 2 PL model, 3 PL model, PCM, etc). This approach emphasizes internal consistency of an individual panelist's judgments (i.e., the same θ value on each item) and the desire for panelists to produce judgments in line with their conceptualized cut score. In an IRT framework, the desire is to arrive at the intended θ cut score for a panelist if the method is implemented correctly. The important extension in Reckase's framework, compared to other methods, is the concept of an intended cut score, which uses the same metric as is used to scale the assessment (i.e., θ metric). This idea of an intended cut score is analogous to an

examinee's *true score* in classical test theory or an examinee's latent ability in IRT, which is a hypothetical construct that is estimated using statistical methods (Reckase, 2006a). That is, the target of the standard setting, the intended cut score, is an unobserved latent variable that the panelist had in mind when they provided their standard setting judgments. Similar to classical test theory, one can conceptualize the estimated cut score as being equal to the intended cut score plus error.

From this perspective, the goal of the person evaluating any standard setting procedure is to determine how each panelist's ratings relate to their cut score and whether these ratings produce the hypothetical cut score that the panelist intended. The important assumption here is that a panelist had their hypothetical preconceived cut score in their minds when they were providing their standard setting judgments and that their standard setting judgments should be in line with the cut score that they were conceptualizing.

Reckase's framework is a major improvement over the other two methods because it allows for potential biases and inconsistencies in panelist cut score estimates to be quantified in the metric of the scale (e.g., the IRT θ -scale or some transformation of the θ -scale) underlying the assessment. That is, his framework addresses the important question of how good the panelists are at estimating the cut scores that they wanted to set on the assessment.

To evaluate a standard setting method, one looks at the amount of statistical bias and imprecision in the standard setting in terms of the hypothetical cut score. In this sense, statistical bias is defined as the difference between the estimated cut score and the hypothetical cut score that a panelist intended to set. In algebraic terms, the concern in Reckase's framework is with:

$$\hat{\theta}_j - \theta_j, \quad (2.12)$$

where θ_j is the true intended cut score for panelist j and $\hat{\theta}_j$ is the estimated cut score from applying the standard setting method. Therefore, in Reckase's framework a high quality standard setting is one in which the amount of bias and imprecision in the cut score estimate is negligible.

2.3.3 *Previous Research Using Reckase's Framework*

Reckase's psychometric theory is relatively new and only two studies that use it have appeared in the research literature. Using simulations based on the Rasch and 3 PL models, the first study (Reckase, 2006a) investigated the potential statistical bias in a single panelist's intended cut score with the Angoff and Bookmark procedures. Results showed the potential impact that rounding item ratings to two decimal places could have on Angoff procedure as well as the potential impact that gaps in the difficulty between items could have in the Bookmark procedure. In general, this first study showed that a panelist's cut score was recovered more accurately with the Angoff method than the Bookmark procedure. The study also suggested that depending on the location of the panelist's desired cut score, the Bookmark method could result in a large amount of statistical bias (Reckase, 2006a).

Responding to Schulz's (2006) criticism of the error models used in his initial study, Reckase (2006b) performed a second investigation of the Angoff procedure in a simulation study using a different set of error models in which the panelist's ratings were regressed toward the mean of the probability scale. Reckase (2006b) showed that this did in fact impact a panelist's estimated cut score in the simulation. He suggested that

additional research was needed using different models for panelist's errors in standard setting.

2.4 Motivation for New Framework

Although the two studies show how to evaluate standard setting methods using Reckase's psychometric theory, neither study provides a clear indication of how to investigate standard setting procedures in operational situations. Consequently, no indices for quantifying potential biases that are directly linked to concept of an intended cut score exist in operational situations.

Indices for evaluating inconsistency in the Angoff method for individual panelists in operational situations do exist. Most of them are discussed in Hurtz and Jones (2009). These indices include indices based on standard errors (Kane, 1987; Hurtz & Hertz, 1999), indices based on absolute deviations from the ICCs (van der Linden, 1982), and the rater balance index (Hurtz & Jones, 2009). Although, it might be possible to adapt and use these indices as measures of how well a panelist's was able to set an intended cut score, the interpretations attached to these indices are not specifically related to the potential biases that might exist when applying the Angoff method. For example, Kane's (1987) index based on standard errors can be used to ascertain fit to an IRT model. It does not, however, quantify the potential impact that the lack of fit has on a panelist's cut score. Currently, no operational indices for evaluating the Bookmark method have been proposed that are directly linked to Reckase's psychometric theory.

Furthermore, in operational situations the concern is often with accurately recovering panelist cut score estimates for a group of panelists. Reckase's psychometric

theory does not however address the question of how to recover panelist estimates for a group of panelists since it was only designed for investigations of individual panelist ratings. An extension of his work is required to perform these investigations. As a result, indices that can be applied to quantify potential biases for a group of panelist's do not exist.

Lastly, Reckase's method requires one to conceptualize how the panelist's ratings relate to the unobserved intended cut score. Developing the conceptualization of this relationship could be challenging. Therefore, it would be useful to create a framework capable of evaluating standard setting methods for either individual or group cut score estimates. This framework should also be capable of clearly linking panelist ratings to intended cut scores.

Therefore, the purpose of this dissertation is to show how Reckase's psychometric theory for evaluating standard setting can be extended to a group of panelists and operational situations. In addition, construct maps (Wilson, 2005) which allow researchers to better conceptualize the relationship of standard setting judgments and intended cut scores are introduced. Chapter 3 explains this new extended framework for evaluating standard setting. Chapter 4 uses the new framework to formulate models and indices for evaluating outcomes of the Bookmark and Angoff standard setting procedures.

2.5 Previous Comparisons of Angoff and Bookmark Procedures

In addition to developing an extended framework for evaluating IRT based standard setting methods, another goal of this dissertation is to compare the performance

of the Angoff and Bookmark methods for operationally setting standards on NAEP in terms of cut score bias. This research is important because despite the widespread use of the Angoff and Bookmark methods in practice, only a small number of empirical comparisons of Angoff derivative methods and Bookmark hybrid procedures have been reported in the research literature.

One such comparison was provided by Buckendahl et al. (2002). Their study compared a modified version of the Bookmark procedure where the items were ordered by observed p -values with the Yes/No procedure (Angoff, 1971; Impara & Plake, 1997) in a K-12 setting in a Midwestern school district. They showed that the two standard setting procedures tended to produce somewhat similar results in terms of the mean cut score estimates, but that the Bookmark procedure had smaller variance in the second and final round of ratings when compared to the Yes/No procedure. Buckendahl et al. (2002) also indicated that there were similar levels of confidence in cut score estimates for panelists who used both procedures.

An issue with this study, however, is that a specific RP value was not used in the Bookmark procedure. Instead, panelists were allowed to apply their own decision rules as to what constituted mastery when indicating their cut score. This makes interpreting the results of comparisons in terms of what one might expect in other comparisons of the Yes/No procedure and Bookmark type procedures difficult since the application of the Bookmark method was far from traditional.

A second comparison of these same standard setting procedures was provided by Davis et al. (2008) on an international licensure exam. Similar to the study by Buckendahl et al. (2002), the overall cut score estimates for the Yes/No procedure and

the modified Bookmark procedure were quite similar. This study differed from Buckendahl et al. (2002) in that panelists participated in both standard setting procedures rather than using two equivalent panels. An RP value of 0.67 was also used with the Bookmark procedure. However, the items in Bookmark were still ordered by the observed *p*-values and each panel participated in Yes/No procedure followed by the Bookmark method. A divergent finding from Buckendahl et al. (2002) was that panelists reported greater confidence when applying the Yes/No procedure than they did in applying the Bookmark procedure.

There were several limitations in the Davis et al. (2008) study, which could explain the similarity of the results of the two methods and limit the generalization of the research. In particular, the Yes/No method was performed first in the two panels which might imply that the similarity of the cut scores for the Yes/No method and Bookmark procedure could be a function of panelists trying to match their Bookmark cut scores to their initial cut scores set with the Yes/No procedure instead of the two methods actually giving similar results in practice. Further, the observed preference of the Yes/No method in the study could be explained by the fact that the panelists invested significant time in learning and performing the method first in comparison to applying the Bookmark method second.

A third comparison of the Bookmark hybrid procedures and Angoff standard setting occurred in the 2005 mathematics pilot study of NAEP (ACT, 2005; Schulz, 2006). In this study, the Mapmark method (Mapmark is explained in greater detail in Chapter 5) was compared to the Angoff method with Mean Estimation across four rounds of ratings. Slight differences between the two methods were observed with the Mapmark

method generally having lower cut scores estimates than the Angoff method. Clear explanations for the differences between the two methods were not given in the study, but the author suggested that one possible explanation for the differences could be from panelists placing their bookmark too early based on perceiving some of the items to be out of order (Schulz, 2006). Reckase (2006a) suggested that a possible explanation was that the Bookmark method can yield negatively biased cut scores due to the way in which the cut scores are estimated and the presence of item difficulty gaps between items.

Schulz (2006) argued in support of the defensibility of the Bookmark standard setting activity based on some concerns related to rater inconsistency in the Angoff method. However, whether or not the rater inconsistency actually explains the observed differences between the two methods and whether the rater inconsistency has the potential to result in bias for the Angoff method was not fully investigated in this study. In addition, potential issues in the cut score estimates in the Bookmark method from item difficulty gaps were also not completely addressed in the study.

Each of these studies are informative because they provide information about how well each of the two most commonly applied standard setting methods compare to each other in practical settings. However, many of the findings reported in these studies could be a function of variations of the Angoff derivative methods and Bookmark-type standard settings that were implemented in the research studies or the context in which the study was conducted. Moreover, the studies do not explicitly consider the potential biases that might be present in applying these standard setting methods in practice or how these biases might impact the cut score estimates. Considering the potential biases in the Angoff and Bookmark procedures, not just whether the cut scores are similar or not, is

important given the high stakes that are often associated with the cut scores. It could be the case, especially in the first two studies that compared the Yes/No method and the modified Bookmark method, that both methods could be biased in similar ways.

Given the widespread use of each of these procedures and the lack of empirical comparisons of the two methods, additional research that looks at the potential biases present in applying the two procedures in operational situations is warranted. If it can be shown empirically that one of the procedures tends to produce greater amounts of potential bias, this could give added support to using one method to set cut scores over another and spawn additional research into how different levels of bias arise when applying the two methods.

Therefore, the empirical illustrations in Chapter 5 of this dissertation will reanalyze the data from the comparison of the Angoff and Bookmark method in the 2005 mathematics pilot study of NAEP (ACT, 2005, Schulz, 2006) for potential biases and inconsistencies. The goal of this reanalysis is to provide a better understanding of how panelists perform the two standard setting methods and how this might impact the cut score estimates on NAEP. This reanalysis could provide greater clarity as to why the Bookmark-type standard setting procedure and the Angoff method performed differently for these data. It might be the case that the differences in the methods are a function of the rater inconsistency as Schulz (2006) suggests or there might be other explanations for the differences that have not yet been identified.

CHAPTER 3

NEW EVALUATION FRAMEWORK

The new framework proposed in this dissertation is an extension of Reckase's psychometric theory for standard setting (Reckase, 2006a) in conjunction with construct maps (Wilson, 2005). This chapter describes both the extensions of Reckase's framework and the concept of construct maps. Additionally, it explains the step-by-step process for evaluating standard setting outcomes.

In Chapter 4 of this dissertation, I demonstrate how the framework that is developed can be used to investigate operational Angoff and Bookmark standard settings for potential biases and inconsistencies. The current chapter illustrates the general framework and how it could be applied to evaluate a hypothetical IRT based booklet standard setting procedure. A separate and general presentation of the framework apart from how the framework can be applied to evaluate standard setting judgments from the Angoff and Bookmark methods is provided in this chapter. This separate presentation is provided to illustrate that the framework that is developed can be applied to evaluate any IRT-based standard setting procedure.

3.1 Extensions of Reckase's Psychometric Framework

Two extensions of Reckase's original psychometric framework are given in this dissertation. The first extension allows the use of Reckase's method for evaluating potential biases and inconsistencies in cut score estimates for a group of panelists, rather than just individual panelists and is based on Reckase's (2006a) original approach for

evaluating standard setting for individual panelists (i.e., Equation 2.12). That is, this extension examines bias at the level of the individual panelist and then aggregates individual panelist ratings in order to obtain overall cut score bias estimates for the group of panelists.

Thus, the potential bias in group cut score estimates can be defined as the difference between the estimated group cut score using the standard setting method and the cut score estimate obtained from combining each of the panelist's hypothetical intended cut scores. This can be represented algebraically as:

$$\frac{\sum_{j=1}^m \hat{\theta}_j}{m} - \frac{\sum_{j=1}^m \theta_j}{m} \quad (3.1)$$

where θ_j is the true intended cut score for panelist j and $\hat{\theta}_j$ is the estimated cut score from applying the standard setting method for panelist j , and m is the number of panelists.

This first extension is based on the critical standard setting assumption that panelist ratings are independent. This assumption is commonly used in cut score and standard error computations (Schulz & Mitzel, 2005). More specifically, one needs to assume that the ratings provided by one panelist are not influenced by the ratings of other panelists or any ratings that a specific panelist made in previous rounds. For example, if panelists are allowed to discuss their ratings it is assumed that the discussion of ratings with other panelists does not create dependencies between the ratings. Unfortunately, this assumption is extremely hard to test in practice since panelists are not typically asked whether their ratings are systematically influenced by other panelists. In addition, there often are a very limited number of observations to quantitatively test for the complex dependencies that might be present.

The second extension in some sense might not be viewed as an extension at all, but rather a demonstration of how the framework can be used to evaluate the potential statistical bias in operational situations. In Reckase's original formulation, he suggested that the key to evaluating any standard setting method was to measure how well a panelist's hypothetical intended cut score was recovered when the method was applied as it would typically be applied in practice. Reckase's (2006a; 2006b) initial demonstrations of his framework consisted of showing the method's use in evaluating Angoff and Bookmark procedure outcomes using a simulation study. He did not indicate how his theory could be used to evaluate standard setting methods in operational situations.

The demonstration of how this framework can be used in operational situations presents additional complications beyond Reckase's (2006a) initial formulation because in operational situations a panelist's hypothetical intended cut score is never known – it can only be estimated. In this sense, just as in many other areas of psychometrics and statistics, the key is coming up with an estimate of the hypothetical construct and then investigating the quality of this estimate. In many areas of statistics and psychometrics, the quality of the parameter estimates is investigated by determining if the parameter estimates are unbiased and precise. In much the same way, cut score estimates can also be examined to see if they are unbiased and precise.

To help guide this evaluation of potential biases and inconsistencies in operational situations, construct maps (Wilson, 2005) are used for providing spatial representations of how panelist ratings are related to the potential cut scores that a panelist intends to set on the assessment. These tabular maps are instrumental in developing statistical models

and indices to quantify the potential biases and inconsistencies that may be present in different standard setting methods.

3.2 Construct Maps

To introduce the concept of construct maps, it is important to recall the IRT models that were introduced in Chapter 2. Each IRT model consists of two sets of unknown quantities. The first set of unknown quantities relates to examinees as reflected by ability parameters. The second set of unknown quantities relates to test items as reflected in item parameters and statistics. A construct map is a spatial representation between the score scale underlying test performance (i.e., the construct) and the examinee and item data on which the IRT models are based. Specifically, construct maps show the relationships between test performance and any quantity that one might derive from an IRT model.

The idea and application of construct maps to provide spatial representations between the underlying latent construct and other components of the measurement model can be traced at least to the work of Wright and Stone (1979) and Wright and Masters (1981) using the Rasch model. However, these authors did not call their graphical output a “construct map”. Instead, the authors discussed the relationships between latent constructs and quantities from measurement models and indicated the usefulness of graphics for depicting these empirical relationships. These early maps included item locations based on item difficulty estimates, a score scale, and histograms representing the distribution of examinees.

Using the basic tenants of these early conceptualizations of the concept, other researchers expanded and adapted the idea of a construct map for the Rasch model. A seminal example is an article by Master et al. (1994) where the general notion of relating the achievement construct with many other quantities that could be derived from the Rasch model is discussed and illustrated.

Since the origin of the idea of a construct map was not given the distinct label “construct map”, spatial representations between underlying latent constructs and quantities from measurement models can be known by several different names in the research literature. These include the terms “Wright map”, “item-person map”, and “variable map” (Bond & Fox, 2007). These latter terms are common in the Rasch literature and describe empirically derived output from the Rasch model showing the relationship between the score scale, item difficulty, and examinee distributions. Versions of construct maps that emphasize relationships between the underlying construct and specific measurement components have also been given distinct names, such as an “item map” (Wang, 2003), Reckase chart (Reckase, 2001), and “domain score chart” (Schulz & Mitzel, 2005). Each of these terms and variations of a construct maps have been used in standard setting.

The term “construct map” is used in this dissertation, as opposed to some of the other terms, because it conveys the idea that the construct can be related to any quantity that can be derived from an IRT model. It also avoids certain ambiguities that might arise with some of the other labels that have been used in the literature. The use of the term in this sense is somewhat similar to the way that Wilson (2005) uses it in his book *Constructing Measures: An Item Response Modeling Approach*, where he discusses how

an underlying construct can be theoretically related to both respondent information and responses to test items. However, the conceptualization in this dissertation is broader than Wilson's (2005) because a Rasch model is not assumed to fit the data. Instead, any IRT model can be assumed including any of the models discussed in Chapter 2. In addition, the specific context for the construct maps is standard setting. Consequently, the quantities included in construct maps include any IRT derived quantity that one might use to determine cut scores. Wilson's (2005) work on construct maps was in the context of instrument and test development and was mainly theoretical. His examples of construct maps did not include many of the quantities that they are conceptualized to contain in this dissertation.

Examples of quantities that are often derived from IRT models and could be included in construct maps are: (1) expected item probabilities (i.e., ICCs), (2) expected performance on content domains (i.e., the proportion-correct true scores for that domain; the TCCs in specific content domains divided by the number of score points), (3) the score scale used for reporting, (4) scale values where individual students are located (i.e., examinee ability estimates), (5) the percentage of students achieving at each score value in the previous year (i.e., the PAC), (6) whole samples of test performance corresponding to particular score scale values, (7) score profiles (i.e., item response vectors for examinees), (8) item locations based on particular response probabilities, and (9) information on demographic subgroup performance. If the tests are vertically scaled, the vertical scale across grades could also be depicted in construct maps. An example of a hypothetical mathematics construct map is provided in Table 3.1.

Table 3.1 shows the score scale that underlies test performance. The score scale in Table 3.1 is a monotonic transformation of the IRT θ -scale and it is displayed in the center of the chart with examinee performance data to the left of the score scale values and item performance data to the right of the score scale values. Data corresponding to a specific θ value is displayed in a single row. The important observation is that most of the quantities that are used to determine cut scores, either as part of the standard setting method itself or as feedback, exist within the construct mapping framework (Table 3.1).

Table 3.1: Hypothetical Mathematics Construct Map

Consequence Data (PAC)	Teacher's Students	Whole Booklets	Score Scale	Item Scores			Domain Scores		
				Item 1	...	Item 50	Number Sense	...	Algebra
...		
14%		K, L	200	.91		.97	.95		.82
19%	Student A	M, N	197	.88		.96	.93		.81
24%	Student B and C	O, P	194	.83		.95	.91		.78
31%	Student D and E	Q, R	191	.77		.94	.88		.74
36%	Student F, G and H	S, T	188	.70		.92	.85		.68
40%	Student I	U, V	185	.63		.91	.82		.64
44%		W, X	182	.55		.89	.79		.59
48%	Student J and K	Y, Z	179	.48		.86	.73		.55
53%		AA, BB	176	.42		.83	.66		.49
59%	Student L	CC, DD	173	.37		.79	.65		.48
...		

Note: The quantities in this table are contrived. The letters in the whole booklets column correspond to booklets that would be presented to a panelist. Similarly, the letters under teacher's students correspond to students in the teacher's classroom.

Construct maps such as the one in Table 3.1 provide a clear indication of what it means to set a cut score at a specific level. For example, if the cut score is set at 185, then this level corresponds to 40% of the students being above the cut score, the performance

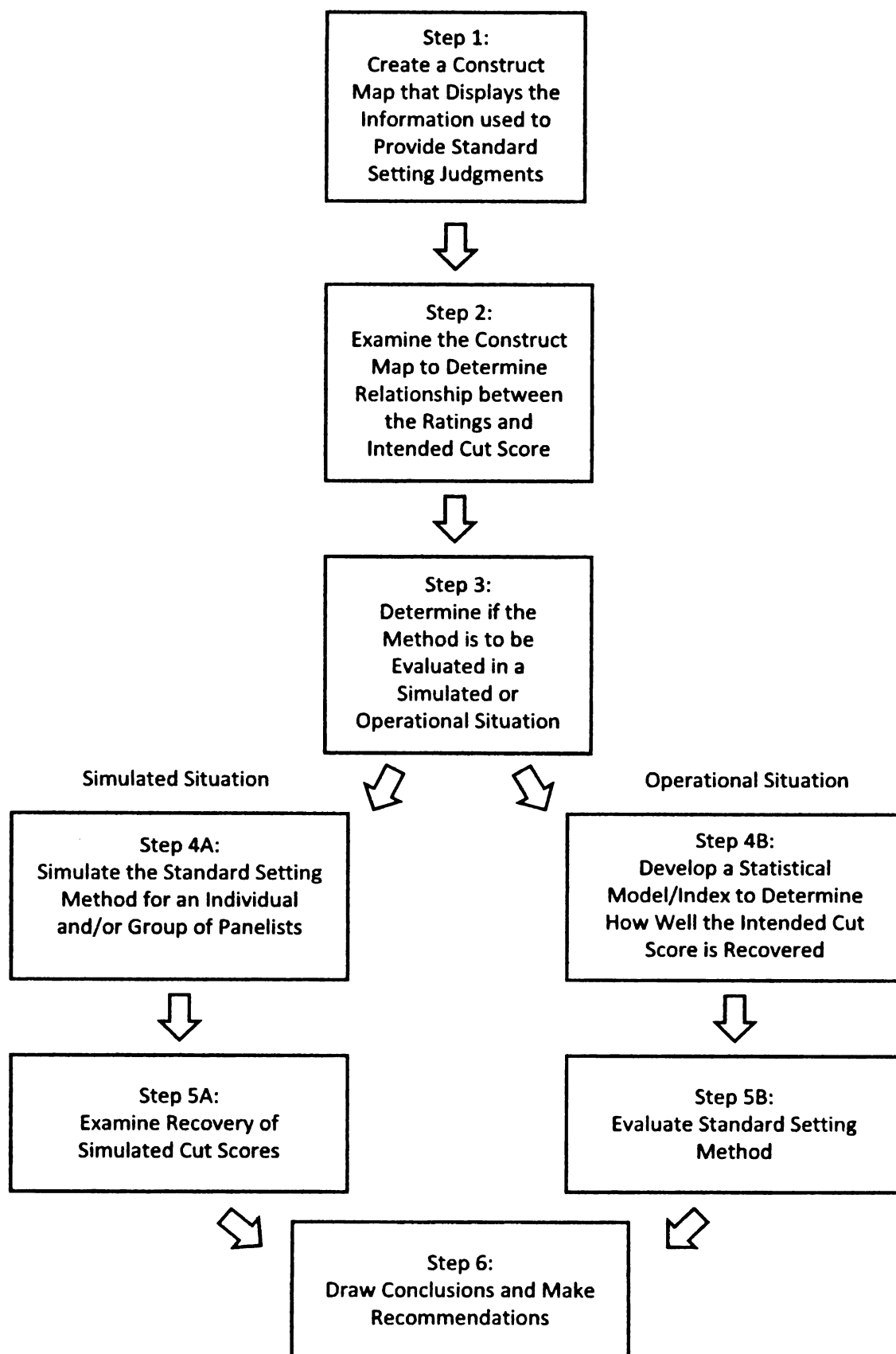
of Student I and the whole booklets U and V, an expected performance on item 1 of 0.63 an expected performance on item 50 of 0.91, an expected performance in algebra of 82 percent and an expected performance in number sense of 64 percent (Table 3.1). Similarly, if the cut score is set at 173, then this level corresponds to 59% of the students being above the cut score, the performance of Student L and the whole booklets CC and DD, an expected performance on item 1 of 0.37, an expected performance on item 50 of 0.79, an expected performance in algebra of 48 percent and an expected performance in number sense of 65 percent (Table 3.1).

Construct maps also provide a clear illustration of how to evaluate any standard setting method because when the standard setting method is working effectively, panelists should be able to set any possible cut score on the score scale by providing a rating that falls into a single row in the construct map. These panelist ratings should correspond to the cut score that the panelist had in mind (their intended cut score) when they provided their standard setting judgments.

3.3 New Comprehensive Evaluation Framework

The new standard setting framework developed in this dissertation draws on the extensions discussed in Section 3.1 and the construct maps presented in Section 3.2. Recall that the first extension is to extend Reckase's psychometric theory to a group of panelists and the second extension is show how Reckase's psychometric theory can be applied in operation situations. Figure 3.1 illustrates the new standard setting evaluation framework.

Figure 3.1: Framework for Evaluating a Standard Setting Procedure



The flow chart in Figure 3.1 shows that the evaluation procedure is a multi-step process with two different paths – one for simulations and the other for operational situations. These two paths are necessary because slightly different approaches are needed for simulated evaluations of standard setting methods compared to operational evaluations of standard setting methods. More importantly, however, all standard setting evaluations can be grouped under a single comprehensive evaluation framework. That is, the new standard setting evaluation approach can be viewed as being comprehensive since both simulation and operational situations can be evaluated under the same framework and the framework can be applied to any IRT-based standard setting method.

To illustrate how this framework might work, a hypothetical booklet based standard setting method is used to illustrate the different steps in the framework. Throughout this section this method will be called the “Whole Booklet” standard setting method. This hypothetical method consists of having panelists select booklets from a set of booklets to represent their cut scores. The average of the booklets that a panelist selects is assumed to be the cut score for an individual panelist. The average of the cut scores for the individual panelists is the cut score for the group of panelists.

3.3.1 Step 1: Create a Construct Map

The first step in performing any standard setting evaluation is to create a construct map that displays the information used to provide the standard setting judgments. The purpose of creating a construct map in the first step is to help the person performing the evaluation to have a clear picture of the relationship between the score scale and the information used to perform the standard setting judgments.

The construct map for the Whole Booklet standard setting method is displayed in Table 3.2. This construct map corresponds to the Whole Booklets column in Table 3.1 since this column corresponds to the stimuli that panelists use to provide their standard setting judgments in the hypothetical Whole Booklet standard setting method.

Table 3.2: Construct Map for Hypothetical Whole Booklet Standard Setting Method

Whole Booklets	Score Scale
	...
K, L	200
M, N	197
O, P	194
Q, R	191
S, T	188
U, V	185
W, X	182
Y, Z	179
AA, BB	176
CC, DD	173
	...

3.3.2 Step 2: Examine the Construct Map and Determine Relationships

After creating the construct map based on the IRT model, the next task is to examine the relationship between the information that is used to perform the ratings and the possible intended cut scores (i.e., the values on the score scale). The goal of this step is to identify whether there are potential issues for panelists in terms of being able to indicate any potential cut score that they could have in mind when providing their standard setting judgments. In general, there are two types of problems that might arise in standard setting that can result in potential biases in cut score estimates: (1) a method design issue in which it is not possible to set cut scores at certain locations along the score scale due to gaps in the score scale from the lack of standard setting stimuli at specific scale locations or (2) the potential for rater inconsistency issues such that

panelists may not be able to provide standard setting judgments that fall into a single row of a construct map. A standard setting method could have either one of both of these potential problems in practice.

To illustrate this step, again consider the Whole Booklet standard setting method. A few important observations can be made from examining the construct map for the Whole Booklet standard setting method in Table 3.2. First, it is apparent that a panelist who understood the Whole Booklet method perfectly could set their cut score at their desired cut score location by selecting the booklets that correspond to the cut score that they want to set in the construct map, as long as there are booklets present at that location. For example, if the intended cut score is 179, booklets Y and Z would be perfect representations of this cut score. These should be the booklets that the panelist selects if they performed the standard setting task correctly.

Second, the construct map clearly shows that there is the possibility for both rater inconsistency and gaps along the score scale to be present when applying the Whole Booklet standard setting method. Rater inconsistency might be present if the panelist who intended to set their cut score at a specific value selected some booklets that are different than the value they intended. For example, the panelist who intended to set their cut score at 179 might select booklets other than booklets Y and Z (e.g., a panelist selects booklets AA, CC, and Z) as representations of their cut score estimate. The selection of the incorrect booklets can lead to potential biases in the panelist's cut score estimate. If the panelist selected booklets AA, CC, and Z, then their cut score estimate would be 176 instead of 179. This means that the cut score that they intended to set would be underestimated by 3 points on the score scale.

The concern of gaps along the score scale could occur when performing Whole Booklet standard setting if the value of the panelist's intended cut score did not have any booklets at this location. For example, if the panelist wanted to set their cut score at 180, there are not any booklets that are displayed at score scale value of 180 in the construct map. This suggests that even if the panelist intended to set their cut score at this value and understood the task of locating booklets that corresponded to the level of 180 they would not be able to perform the standard setting task correctly. Both of these situations, either separately or combined, indicate that depending on the intended cut score and the ratings provided by the panelist there is the potential for bias in the standard setting task. Whether the method would produce bias in a given situation is an empirical question that is investigated in subsequent steps in the framework.

3.3.3 Step 3: Determine How Method is to be Evaluated

The next step is to determine whether the method is to be evaluated in a simulated or operational situation. This determination is important because different approaches are used to conduct the evaluation in each case. In a simulated evaluation, the person performing the evaluation knows the values of true intended cut scores for an individual panelist or a group of panelists. Since these quantities are known, one can proceed to look at how well these values would be recovered under the conditions specified in the simulation. If the evaluation is instead to be performed in an operational standard setting, a different set of procedures is needed to conduct the evaluation since one does not know the true value of the intended cut score; one only knows the ratings provided by the panelist and the estimated cut score from these ratings. Hence, the important question

becomes given the ratings provided by the panelist, how well could the intended cut score be recovered from these ratings?

Since the methods for evaluating a standard setting diverge somewhat depending on whether the method is evaluated in a simulated or operational situation, both situations are discussed separately below. The steps to evaluate a method in a simulated situation are discussed followed by the steps to evaluate a standard setting method operationally. The last step in the framework draws together these two divergent paths.

3.3.4 Step 4A: Simulate the Standard Setting Method

One might choose to perform a simulated investigation of a standard setting procedure if they are interested in investigating how a standard setting method would perform in various situations or in different contexts without actually conducting an operational standard setting. This can shed important insight into the functioning of a standard setting method in a situation similar to what might be encountered in practice without incurring the substantial costs of operationally setting standards. For example, one might choose to simulate a standard setting method before operationally implementing a standard setting method to see how well the judgments of a hypothetical group of raters would be recovered in a specific context.

The advantage of this approach is that investigator has control of the variables that might impact the cut score judgments and the true intended cut score for the panelists or group of panelists are known in the investigation. This allows the recovery of the estimated cut scores to be directly compared to the true intended cut scores. The disadvantage of using simulations is that the factors manipulated in the simulation are the

only factors that can impact standard setting judgments. The factors that are manipulated may or may not be representative of the factors or how they would function in an operational standard setting situation.

There are many different possibilities for how to simulate and evaluate the standard setting method. Some important considerations in developing a simulated evaluation are the following:

- 1) Is the standard setting method going to be simulated for an individual panelist and/or group of panelists?
- 2) What distribution or distributions of panelist cut scores should be considered?
- 3) Do the panelists perform the standard setting method perfectly or do they make errors? If they make errors, what model or models should be used for the panelists' errors?
- 4) Does the standard setting process consist of multiple rounds? If so, what model should be used to simulate the ratings and interactions of the panelists across rounds? Do the ratings and distributions of panelists change over rounds?

The four questions above are just a sampling of the questions that a person evaluating a standard setting might consider. The questions asked and the simulation designed could include other questions and be considerably more complicated.

An example of a potential simulated evaluation for the Whole Booklet standard setting method could be to evaluate the method for a group of twenty panelists in a single round assuming that each of the panelists are able to locate the booklets closest to their simulated cut score estimate with the distribution of cut score estimates for the

group of panelists assumed to be standard normal. It is this step of simulating and evaluating the recovery of the panelist hypothetical cut scores that gives power to Reckase's (2006a; 2006b) initial psychometric theory for evaluating standard setting since this framework allows researchers and practitioners to investigate the functioning of standard setting methods in any hypothetical situation.

In his initial work, Reckase (2006a; 2006b) conducted simulated evaluations of the Bookmark and Angoff procedure for an individual panelist assuming that the standard setting process consisted of a single round with two distinct models for panelist errors. Specifically, Reckase used a beta distribution (Reckase, 2006a) or regression of the ratings toward the mean of the probability scale (Reckase, 2006b) and a range of cut scores from $\theta = -3$ to $\theta = 3$.

3.3.5 Step 5A: Examine Recovery of Simulated Cut Scores

The last step in evaluating a standard setting procedure using simulations is to check how well the simulated cut scores are recovered. Typically, parameter recovery in a simulated situation is assessed by examining the difference between estimated value(s) and the simulated value(s). It is the difference between the estimated value and the simulated value that indicates how well the method works under the conditions specified in the simulation. When the difference between the simulated value and the estimated value in the simulation is not zero this means that there is the potential for standard setting method to be biased under those conditions. Often, the concern is with the bias of an individual panelist, which is Equation 2.12 from Reckase's original psychometric

theory. The concern might also be with the bias at the group level, which is the extension of Reckase's psychometric theory outlined in Equation 3.1.

In the simulated evaluation of the Whole Booklet standard setting method discussed in the previous section, Equation 2.12 could be examined for the twenty individual panelists and Equation 3.1 could be examined for the group of panelists. The desire in both cases is for these statistics to be zero since this would indicate perfect parameter recovery and an unbiased standard setting method under the conditions specified in the simulation. In Reckase (2006a; 2006b), the individual panelist ratings from $\theta = -3$ to $\theta = 3$ for the Bookmark and Angoff standard setting methods were evaluated using Equation 2.12. This shows that Reckase's (2006a; 2006b) initial evaluation framework and investigations are a subset of the complete framework that is being developed in this dissertation.

3.3.6 Step 4B: Develop a Statistical Model or Index to Evaluate Method

Since the true intended cut score is not known in operational situations and can only be estimated, a different approach to evaluating the standard setting method is needed in operational situations. In this case, one needs to examine the ratings provided by the panelist to determine how well the ratings represent the potential intended cut scores. This inspection requires the development of statistical models or indices to evaluate the method. The development of these models or indices draws directly from the construct map in the second step of the framework. In this second step, the construct map is used to ascertain whether it is possible for a panelist to provide ratings that are consistent with any hypothetical intended cut score that they want to set and whether

there are potential threats to providing these ratings when performing the standard setting procedure. The indices or statistical models that are generated to evaluate the standard setting procedure should be measures of the extent to which the threats of potential gaps along the score scale or panelist inconsistency could impact the cut score estimates. The development of these indices always starts at the individual panelist level and then extends the indices to a group of panelists if one is interested in investigating the potential issues at the group level.

For example, an actual set of ratings provided by a panelist for the hypothetical standard setting method could be selecting booklets U, S, M, and P as representations of the cut score that they want to set. The ratings provided by this panelist are underlined and italicized in Table 3.3 below.

Table 3.3: Individual Panelist Ratings for Hypothetical Booklet Standard Setting

Whole Booklets	Score Scale
	...
K, L	200
<u>M</u> , N	197
O, <u>P</u>	194
Q, R	191
<u>S</u> , T	188
<u>U</u> , V	185
W, X	182
Y, Z	179
AA, BB	176
CC, DD	173
	...

Clearly, the panelist who performed this standard setting did not identify booklets in the construct map that fall into a single row. The task then is coming up with a measure of how good the ratings provided by the panelist are in representing their estimated cut score. In this case, the panelist's estimated cut score would be 190.25 since

this is the average of 197, 191, 188, and 185. There are many indices that can be developed in this step to quantify the quality of the cut score estimates. Two potential indices for measuring the quality of the cut score estimates for this standard setting procedure are the average and absolute residuals for that panelist. These two indices are defined as:

$$\frac{\sum_{i=1}^r \hat{e}_i}{r} \quad (3.2)$$

and,
$$\frac{\sum_{i=1}^r |\hat{e}_i|}{r} \quad (3.3)$$

with
$$\hat{e}_i = \text{ScoreScale}_i - \frac{\sum_{i=1}^r \text{ScoreScale}_i}{r} \quad (3.4)$$

and,

$$|\hat{e}_i| = \left| \text{ScoreScale}_i - \frac{\sum_{i=1}^r \text{ScoreScale}_i}{r} \right|, \quad (3.5)$$

where ScoreScale_i is the score scale value for rating i , and r is the number of ratings. Equation 3.2 and 3.3 provide indications of the magnitude of the impact of the panelist's inconsistency. Equation 3.2 could be used to determine whether the errors cancel out over the ratings, while Equation 3.3 could be used provide an indication of the extent of the absolute errors made by the panelist. These indices could also be averaged over the panelists to provide measures of the average errors across all panelists and the average magnitude of the absolute errors of the group of panelists. Other indices besides these are

possible, but the key is developing models or indices that can be used to give clear indications of quality of the cut score estimate as well as the potential for biases present in ratings.

At this point it is important to point out that in evaluating an operational standard setting method one needs to make an assumption about the cut score estimate provided by the panelists in order to perform the evaluation. The critical assumption used in this framework is that the estimated cut score in operational situations should be viewed as a representation of the cut score that the panelist intended to set. The goal of the evaluation is then is to determine how good panelists are at estimating their intended cut score. Some limitations of this assumption and framework are presented in the discussion in Chapter 6.

3.3.7 Step 5B: Evaluate Standard Setting Method

The last step in evaluating an operational standard setting procedure is to actually calculate the indices and use the models developed in Step 4B to determine the quality of the cut score estimates. This step is straightforward and is just an application of the indices and models to the ratings that the panelists provided. For the hypothetical standard setting method based on the whole booklets considered throughout this section, the indices in Equations 3.2 and 3.3 would be applied to each of the individual panelist ratings. These indices could also be aggregated and tabulated at the group level.

3.3.8 Step 6: Draw Conclusions and Make Recommendations

The last step in the framework is to draw conclusions and make recommendations based on the evaluations that have been carried out. In a simulated situation, these conclusions might be that the method either works quite well or not so well in the conditions specified in the simulation. These simulated investigations could be very valuable before using a procedure operationally because they could provide the standard setter with essential information about how the standard setting method might be expected to perform. If the standard setting procedure has the potential to result in large amounts of statistical bias either at the individual or group level this might prevent one from using the method to set standards operationally.

In operational situations, the framework can be used to draw important conclusions about how well the panelists were able to perform the standard setting task. This information could be very helpful to a policy board as they deliberate and make a final decision about whether or not to adopt the panelists' recommended cut scores. If the indices and models from the operational standard setting identify potential issues with the standard setting procedure, this could be used as justification to change the cut scores or to use a different standard setting procedure in the future. An additional advantage of this framework might be the ability to apply the framework after each individual round of a standard setting process and to use the information from the statistical indices as feedback in the next round to help panelists improve their standard setting ratings.

CHAPTER 4

INDICES FOR EVALUATING ANGOFF AND BOOKMARK

This chapter shows how the new comprehensive standard setting evaluation framework can be used to construct indices to evaluate operational Angoff (the Angoff method with Mean estimation) and Bookmark standard setting procedures. Specifically, separate indices are constructed for investigations of the cut scores estimates for an individual panelist or group of panelists using steps 1 through 4B in the comprehensive framework developed in Chapter 3. The application of these indices to evaluate operational standard settings, steps 5B and 6, is provided in Chapter 5. The reason for developing indices to quantify the potential biases for these two standard setting methods is that they are among the most commonly applied standard setting methods and indices to quantify the potential biases in operational situations in relationship to intended cut scores do not exist for these two procedures.

To facilitate their comparison for both standard setting methods, the indices that are developed are placed onto common scales that are easy to interpret. Specifically, one set of indices is developed in the θ -metric (or scale score metric) and another set of indices is developed to quantify the potential changes in the PAC (percent above the cut score) metric for that cut score. An application of these new indices to answer the research questions of Section 1.4 is provided in Chapter 5 using data from previous NAEP standard settings.

4.1 Data to Illustrate New Evaluation Methods

Throughout this chapter, a set of hypothetical data will be used to create the construct maps and illustrate potential issues that might arise when using the Angoff method with Mean Estimation and Bookmark method. For didactic purposes, it is assumed that the standard setting is conducted on a nine-item test. Although a nine-item test is significantly shorter than test lengths for many assessments delivered operationally, the use of a nine-item test allows the examples to be presented in a clear and concise format. In practice, tests typically contain dozens of items.

The nine-item test has six multiple-choice items that are calibrated using the 3PL model, two short-answer questions that are calibrated using the 2PL model, and a constructed-response item with three score points (0, 1, and 2) that is calibrated using the GPCM. Each IRT model is described in Chapter 2. Item parameters for the nine items are given in Table 4.1.

Table 4.1: Item Parameters for the Nine Items Used in the Didactic Examples

Item Number	Item Type	a	b	c	d_1	d_2
1	Multiple choice	0.741	-0.694	0.187		
2	Multiple choice	0.915	-0.873	0.197		
3	Multiple choice	1.670	0.269	0.258		
4	Multiple choice	0.468	-1.311	0.156		
5	Multiple choice	0.961	0.681	0.126		
6	Multiple choice	1.436	1.024	0.130		
7	Short answer	2.084	0.836	0.000		
8	Short answer	0.871	-0.228	0.000		
9	Constructed response	0.728	-0.985	0.000	-0.741	0.741

The item parameters and item types in Table 4.1 are similar to those from previous NAEP administrations in that multiple-choice, short-answer, and constructed-response items fit by the 3PL, 2PL, and generalized partial credit model are displayed in Table 4.1. Similar to NAEP, the nine-item test has more multiple-choice items than the other two item types. The item parameters are representative of the item parameters of actual NAEP items.

A closer examination shows that some items are more discriminating than others and that the items are evenly distributed along the difficulty scale. All the pseudo-guessing parameters have values that are less than 0.300 indicating low levels of guessing.

4.2 Methods for Evaluating Angoff Method Outcomes

Panelists that use the Angoff method with Mean Estimation are asked to indicate the expected probability of the MCE answering the dichotomous items correctly and the mean expected performance of the MCE on the polytomous items. In order to determine the cut score estimate for an individual panelist on the θ -scale, the items scores for a MCE are summed and converted to θ -scale using the relationships between the true scores and θ values expressed in the TCC. The cut score for a group of panelists is usually either the mean or median of the cut score estimates of the individual panelists.

Using the framework outlined in Chapter 3, indices for evaluating the Angoff method with Mean Estimation are developed. First, a sample construct map is shown in Table 4.2 for the Angoff method with Mean Estimation. For simplicity, it is assumed that the panelist is asked to perform the Angoff method with Mean Estimation on the nine

item test presented in Section 4.1. The construct map for select θ values from $\theta = -3$ to $\theta = 3$ for these nine items is displayed in Table 4.2. (Note that the construct map in Table 4.2 could be expanded to other θ locations).

Table 4.2: Construct Map for the Angoff Method

θ	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
3.000	0.992	0.998	1.000	0.974	0.981	0.993	1.000	0.992	1.991
2.750	0.990	0.997	0.999	0.968	0.971	0.987	0.999	0.988	1.988
2.500	0.986	0.996	0.999	0.961	0.957	0.977	0.997	0.983	1.984
2.250	0.981	0.994	0.997	0.953	0.937	0.958	0.993	0.975	1.979
2.000	0.974	0.991	0.995	0.943	0.909	0.926	0.984	0.964	1.972
1.750	0.964	0.987	0.989	0.932	0.870	0.874	0.962	0.949	1.962
1.500	0.952	0.980	0.978	0.919	0.818	0.793	0.913	0.928	1.949
1.250	0.935	0.972	0.957	0.903	0.753	0.682	0.813	0.899	1.930
1.000	0.914	0.959	0.917	0.884	0.674	0.552	0.641	0.860	1.905
0.750	0.887	0.940	0.849	0.863	0.588	0.425	0.424	0.810	1.869
0.500	0.852	0.915	0.746	0.838	0.499	0.319	0.233	0.746	1.818
0.250	0.810	0.881	0.619	0.811	0.415	0.244	0.111	0.670	1.748
0.000	0.761	0.836	0.494	0.780	0.342	0.196	0.049	0.584	1.653
-0.250	0.704	0.779	0.396	0.746	0.283	0.167	0.021	0.492	1.528
-0.500	0.643	0.712	0.333	0.710	0.237	0.151	0.009	0.401	1.371
-0.750	0.579	0.637	0.297	0.671	0.203	0.141	0.004	0.316	1.187
-1.000	0.516	0.559	0.278	0.630	0.179	0.136	0.001	0.242	0.988
-1.250	0.457	0.484	0.268	0.588	0.162	0.133	0.001	0.180	0.789
-1.500	0.403	0.417	0.263	0.546	0.150	0.132	0.000	0.132	0.608
-1.750	0.357	0.360	0.260	0.505	0.142	0.131	0.000	0.095	0.455
-2.000	0.319	0.316	0.259	0.465	0.137	0.131	0.000	0.068	0.334
-2.250	0.287	0.281	0.259	0.427	0.133	0.130	0.000	0.048	0.242
-2.500	0.263	0.256	0.258	0.392	0.131	0.130	0.000	0.033	0.175
-2.750	0.244	0.238	0.258	0.360	0.129	0.130	0.000	0.023	0.126
-3.000	0.229	0.225	0.258	0.331	0.128	0.130	0.000	0.016	0.092

The construct map in Table 4.2 shows the expected performance for the nine items in rows that correspond to θ values from the IRT model. The values in Table 4.2 for each item are simply the values of the ICC for that item at that θ value. For example, 0.403 is the value of item characteristic curve for item 1 at $\theta = -1.500$. Table 4.2 is an example of a Reckase chart (Reckase, 2001), which is a common feedback mechanism on NAEP.

From Table 4.2 the potential issues that might arise when applying the Angoff method with Mean Estimation can be determined. Notice that even though Table 4.2 has been shortened for illustrative purposes that the relationships that exist in Table 4.2 can be extended to any possible θ location since the columns under each item are just the values of the ICC at that θ location. This means that the Angoff method does not suffer from the problem of gaps along the score scale since it is possible for a panelist to indicate any possible intended cut score by providing standard setting judgments that would be equal to the value of the ICCs of the items at the θ location represented by their intended cut score.

That is, in an ideal Angoff standard setting the item ratings provided by the panelist would fall into a single row corresponding to one θ value, which is their intended cut score. For example, to obtain an ideal standard setting for a panelist who conceptualized their cut score to be at $\theta = -1.500$, the panelist should give item ratings of 0.403, 0.417, 0.263, 0.546, 0.150, 0.132, 0.000, 0.132, 0.608 on the nine items. This row is in bold italics in Table 4.2.

If a panelist does not provide these ratings, this indicates that the panelist is not consistent in their item ratings when the assumptions for the IRT model hold (e.g., the four IRT assumptions in section 2.3.1; monotonicity, statistical independence, functional form, and population and parameter invariance). For example, if the panelist provided ratings of 0.403, 0.559, 0.259, 0.505, 0.162, 0.167, 0.111, 0.095, and 0.242 on the nine items, respectively (see Table 4.3), and the cut score that they intended to set is $\theta = -1.500$, the panelist has not performed the rating task consistently. Therefore, one would conclude that the panelist is not inline with their intended cut score. This lack of

consistency can lead to potential biases in a panelist's intended cut score and is an example of the problem of the rater inconsistency that can occur with some standard setting methods that was discussed in Chapter 3.

An example of a hypothetically inconsistent rater is shown in Table 4.3.

Table 4.3: Hypothetically Inconsistent Rater

θ	item 1	item 2	Item 3	item 4	item 5	Item 6	item 7	item 8	Item 9
3.000	0.992	0.998	1.000	0.974	0.981	0.993	1.000	0.992	1.991
2.750	0.990	0.997	0.999	0.968	0.971	0.987	0.999	0.988	1.988
2.500	0.986	0.996	0.999	0.961	0.957	0.977	0.997	0.983	1.984
2.250	0.981	0.994	0.997	0.953	0.937	0.958	0.993	0.975	1.979
2.000	0.974	0.991	0.995	0.943	0.909	0.926	0.984	0.964	1.972
1.750	0.964	0.987	0.989	0.932	0.870	0.874	0.962	0.949	1.962
1.500	0.952	0.980	0.978	0.919	0.818	0.793	0.913	0.928	1.949
1.250	0.935	0.972	0.957	0.903	0.753	0.682	0.813	0.899	1.930
1.000	0.914	0.959	0.917	0.884	0.674	0.552	0.641	0.860	1.905
0.750	0.887	0.940	0.849	0.863	0.588	0.425	0.424	0.810	1.869
0.500	0.852	0.915	0.746	0.838	0.499	0.319	0.233	0.746	1.818
0.250	0.810	0.881	0.619	0.811	0.415	0.244	0.111	0.670	1.748
0.000	0.761	0.836	0.494	0.780	0.342	0.196	0.049	0.584	1.653
-0.250	0.704	0.779	0.396	0.746	0.283	0.167	0.021	0.492	1.528
-0.500	0.643	0.712	0.333	0.710	0.237	0.151	0.009	0.401	1.371
-0.750	0.579	0.637	0.297	0.671	0.203	0.141	0.004	0.316	1.187
-1.000	0.516	0.559	0.278	0.630	0.179	0.136	0.001	0.242	0.988
-1.250	0.457	0.484	0.268	0.588	0.162	0.133	0.001	0.180	0.789
-1.500	0.403	0.417	0.263	0.546	0.150	0.132	0.000	0.132	0.608
-1.750	0.357	0.360	0.260	0.505	0.142	0.131	0.000	0.095	0.455
-2.000	0.319	0.316	0.259	0.465	0.137	0.131	0.000	0.068	0.334
-2.250	0.287	0.281	0.259	0.427	0.133	0.130	0.000	0.048	0.242
-2.500	0.263	0.256	0.258	0.392	0.131	0.130	0.000	0.033	0.175
-2.750	0.244	0.238	0.258	0.360	0.129	0.130	0.000	0.023	0.126
-3.000	0.229	0.225	0.258	0.331	0.128	0.130	0.000	0.016	0.092

If the intended cut score for this rater is $\theta = -1.500$, then this panelist has not performed the standard setting task consistently since the ratings are scattered at various θ values in the construct map and are not all in the row that corresponds to $\theta = -1.500$ (Table 4.3). In an operational standard setting, the differences or absolute differences of the item ratings from the estimated cut score can be used to ascertain how well the

panelist is able to perform the standard setting task. For example, the difference between the item rating and overall intended cut score for item 1 is 0 θ -units, for item 2 is 0.500 θ -units, for item 3 is -0.500 θ -units, and so on.

4.2.1 Indices for Evaluating Angoff Method Outcomes

The differences and absolute differences between the item ratings for each item and the overall estimated cut score across all of the items can be formulated as residuals and absolute residuals, respectively, and can be used to evaluate the Angoff method in operational situations. These residuals can be represented as:

$$\hat{e}_{ijkl} = \hat{\theta}_{ijkl} - \hat{\theta}_{cjl}, \quad (4.1)$$

where \hat{e}_{ijkl} is the estimated residual for item i for panelist j for performance level k for round l , $\hat{\theta}_{ijkl}$ is the observed score scale rating for item i for panelist j for performance level k for round l derived from the IRT model, $\hat{\theta}_{cjl}$ and is the estimated cut score for the performance level. The absolute residuals are represented as:

$$|\hat{e}_{ijkl}| = |\hat{\theta}_{ijkl} - \hat{\theta}_{cjl}|, \quad (4.2)$$

with the symbols and subscripts retaining the same meaning as Equation 4.1. The use of residuals to evaluate properties of statistical and psychometric models in general and aspects of IRT in particular are quite common (see Hambleton, et al., 1991, for example). These residuals or absolute residuals can be used in statistical models to evaluate Angoff standard setting. These models can be formulated as:

$$\hat{e}_{ijkl} = \mathbf{x}_{ijkl}\boldsymbol{\beta} + c_i + c_j + c_k + c_l + c_{ij} + c_{ik} + c_{il} + c_{jk} + c_{jl} + c_{kl} + c_{ijk} + c_{ikl} + c_{jkl} + u_{ijkl}, \quad (4.3)$$

where \hat{e}_{ijkl} is the estimated residual for item i for panelist j for performance category k for round l , x_{ijkl} is a matrix of covariates, β is a vector of regression coefficients, the c 's are unobserved fixed effects, typically estimated using dummy variables, and u_{ijkl} is a random error. For different research questions of interest, different fixed effects and terms would be estimated in Equation 4.1, while others could be excluded and could become part of the error term.

This model is general and also allows for other covariates, such as demographic or item characteristics, to be added to the analyses and tested for significance depending on the fixed effects being estimated. The fixed effects that are most often of interest are the c_{jkl} effects, which provide estimates of the average residuals for panelists in each round for each performance category, and the c_{kl} effects which provide indicators of the average residuals across panelists within rounds for each performance category. The c_{jkl} effects with the residuals as dependent variables are comparable to Reckase's (2006a) indices for evaluating the bias of individual panelists in a simulated situation. The only difference is that these effects are estimated from operational standard setting data and assume that a panelist's estimated cut score is their intended cut score. There is no analog to the c_{kl} effects in Reckase's (2006a) original formulation of his psychometric theory for standard setting.

The desire would be for these coefficients to be zero since this indicates that there is not the potential for bias in the panelists' estimated cut scores. Other important hypotheses could be tested by assessing the statistical significance of the other variables.

Models could also be formulated with the absolute residuals as the dependent variable, which would allow for hypotheses about inconsistency to be tested.

To facilitate a comparison of the potential biases in the Angoff procedure with Mean Estimation to those that are developed for the Bookmark procedure, a set of comparable indices are needed. Therefore, two sets of indices are developed. The first set is in the θ -metric or the metric of the score scale and focuses on the potential biases of an individual panelist or group of panelists in these metrics. These indices are formulated by simply examining the absolute magnitude of the c_{jkl} and c_{kl} effects in the models in Equation 4.3 which uses the residuals as a dependent variable with no covariates. These effects provide unadjusted estimates of the potential biases in an individual panelist and group of panelists' ratings, respectively.

The second set of indices focuses on the practical impact of the biases on the percentage of students classified above the cut scores – the PAC. To formulate these indices one needs to first add the c_{jkl} or c_{kl} effect from the models with the residuals as dependent variables to the estimated cut score at the panelist or group level, respectively. That is, one needs to compute

$$\theta_{c_{jkl}} + c_{jkl}, \quad (4.4)$$

and

$$\theta_{c_{\bullet kl}} + c_{kl}, \quad (4.5)$$

where \bullet is used to denote aggregation over that factor. In Equation 4.5, the aggregation is over the panelists and $\theta_{c_{\bullet kl}}$ represents the cut score estimate for the group of panelists for performance level k in round l .

Then, one needs to calculate

$$F(\theta_{cjk} + c_{jk}) - F(\theta_{cjk}) \quad (4.6)$$

or

$$F(\theta_{c \bullet k} + c_{k}) - F(\theta_{c \bullet k}), \quad (4.7)$$

at the panelist or group level, respectively, where $F(x)$ gives the percentage of students at or above the cut score x (i.e., the PAC). The indices in Equations 4.6 and 4.7 can be used to determine potential changes in the PAC given panelist inconsistency in Angoff standard setting. The desire for these indices would be that they should be as close to zero as possible since this indicates that the panelists were very consistent in providing their judgments in the Angoff procedure. Additionally, this would indicate that the potential for large standard setting biases to have a practically significant impact on the PAC is minimal.

4.3 Methods for Evaluating Bookmark Method Outcomes

Panelists that use the Bookmark method are asked to move through a booklet of items ordered from (i.e., an OIB) easiest to hardest based on a RP criterion. Recall, that the RP criterion is a probability level that is used to order the items in the OIB. Additionally, in the Bookmark procedure, each panelist moves through the OIB asking themselves whether or not the MCE should be able to answer that item at that score level or higher with a probability greater than or equal to the RP criterion. If the answer to this question is yes, then the panelist moves to the next item. If the answer to this question is no, then the panelist places a mark in their booklet between this item and the item preceding it. Oftentimes, this is represented in practice by placing the bookmark in the form of a post-it note on the item when they answer no to the above question. The cut score for the panelist is determined from the θ location of the item directly preceding the

bookmark. The overall cut score for the group of panelist is the mean or median of the cut scores of all of the panelists.

To illustrate how to create a construct map for the Bookmark method, the nine items in Section 4.1 are used to create a construct map assuming that the RP criterion is 0.67. When the RP criterion is 0.67 this means that there is a 67 percent chance of obtaining a score at that level or higher on that test item. The first step in creating the construct map for the Bookmark method is to locate the θ value where the probability of obtaining a score at that level or higher is equal to 67 percent on each item. For dichotomous items, this is simply the location where the probability of getting a correct response is 67 percent. For polytomous items, such as item 9, the item is placed in the booklet multiple times, one time for each score point above zero that an examinee can obtain on the item. For item 9, the item is placed into the booklet two times; once when the probability of obtaining a score of 1 or higher is equal to 67 percent (Item 9_1 in the chart below) and once when the probability of obtaining a score of 2 or higher is equal to 67 percent probability (Item 9_2 in the chart below).

The order of items in the ordered item booklet can be determined by locating the θ values from the IRT models in Table 4.2 in which the value of the expected probability of correct response equals 0.67 for dichotomous items. For example, in Table 4.2, the item 1 column would be moved up until an RP of 0.67 is located. The θ value corresponding to 0.67 is the Bookmark location of this item when the RP criterion is 0.67. This process would then be repeated for each item in Table 4.2. For item 9, the chart in Table 4.2 should be expanded to include the probabilities of obtaining a score of one or higher (i.e., 9_1) and a score of two or higher (i.e., 9_2) and the θ location where

the probability is equal to 0.67 in each of these cases is determined. The Bookmark procedure construct map for the nine items is depicted in Table 4.4.

Table 4.4: Construct Map for the Bookmark method with a RP of 0.67

Bookmark Location	Item
$+\infty$	After Item 6
1.226	Item 6
1.036	Item 7
0.987	Item 5
0.347	Item 3
0.250	Item 8
-0.298	Item 9_2
-0.392	Item 1
-0.642	Item 2
-0.754	Item 4
-0.797	Item 9_1
$-\infty$	Before Item 9_1

Table 4.4 shows that there are only twelve possible cut scores ($-\infty$, -0.797, -0.754, -0.642, -0.392, -0.298, 0.250, 0.347, 0.987, 1.036, 1.226, and $+\infty$) on the assessment since the cut score for an individual panelist is determined from the θ location in the OIB that precedes the bookmark (Table 4.4). For example, a panelist indicates $-\infty$ if they place their bookmark before item 9_1 (e.g., the location of obtaining a score of 1 or higher on item 9) in the booklet, -0.797 if the bookmark is between item 9_1 and item 4, and so on. The construct map shows that there are item difficulty gaps between the locations where the possible cut scores can be placed (Table 4.4). For example, between the score values of -0.642 and -0.754 there is an item difficulty gap since the cut score cannot be set there (i.e., -0.725) even if this is where the panelist's intended cut score is located. This is an example of a method design concern of gaps along the score scale in a standard setting method. This suggests that to evaluate the Bookmark standard setting

outcomes, one should examine the item difficulty gaps in the region where the panelist places their bookmark. A simple set of statistical indices can be developed for quantifying potential biases in the operational Bookmark standard setting procedure using these item difficulty gaps.

4.3.1 *Indices for Evaluating Bookmark Method Outcomes*

To develop these indices, consider that when a panelist bookmarks an item they indicate that their cut score is some place between the θ values associated with the item before and after the bookmark. That is, even if a panelist understands the method perfectly and performs the method correctly there is indeterminacy as to location of their intended cut score. The cut score is located somewhere between the θ value associated with the item before and after the bookmark. If the gap between the items where the panelist wants to set their cut score is large, then there is the potential for inaccuracies in standard setting even if a panelist fully understands the standard setting task. If the gap is small, then the potential impact is also small. The smallest value that the panelist could indicate for their cut score is the item before the bookmark and the largest value is the item is the item after the bookmark. Define these two quantities as:

$$\theta_S \tag{4.8}$$

and $\theta_L,$ (4.9)

respectively, where θ_S represents the smallest θ value and θ_L represents the largest θ value. For an individual panelist the maximum potential bias in the θ metric is defined as

$$\theta_L - \theta_S. \tag{4.10}$$

For a group of panelists, one needs to compute the cut score for the test based on all the panelists' θ_S and θ_L values, respectively. These quantities can be represented algebraically as:

$$\frac{\sum_{i=1}^m \theta_S}{m} \quad (4.11)$$

and $\frac{\sum_{i=1}^m \theta_L}{m}, \quad (4.12)$

when the average is used to compute the overall cut score and m is the number of panelists. The cut score could also be computed using the median, which could change the cut score estimates from those in Equations 4.11 and 4.12. Equation 4.11 and 4.12 give the range of possible cut scores assuming that the panelists understand the Bookmark task and have performed it correctly. Subtracting Equation 4.11 from Equation 4.12 yields the maximum potential bias in the group cut score in the metric of the score scale. This can be represented algebraically as:

$$\frac{\sum_{i=1}^m \theta_L}{m} - \frac{\sum_{i=1}^m \theta_S}{m}. \quad (4.13).$$

The practical impact of these potential biases can be determined by finding the difference in the PAC based on the largest and smallest cut score that a panelist could be conceptualizing. At the individual panelist level, this is represented as:

$$F(\theta_L) - F(\theta_S), \quad (4.14)$$

and at the group level this can be written as:

$$F\left(\frac{\sum_{i=1}^m \theta_L}{m}\right) - F\left(\frac{\sum_{i=1}^m \theta_S}{m}\right), \quad (4.15)$$

where $F(x)$ again returns the percentage of students at or above the cut score x .

These simple indices quantify the potential change in PAC based on the maximum and minimum possible values that panelists could be conceptualizing when applying the Bookmark method. The goal is for each of these indices to be small and close to zero since this indicates that the potential for changes in the cut score and PAC from the item difficulty gaps in the Bookmark procedure is minimal.

In the θ or scale score metric, the indices in Equations 4.10 and 4.13 can be directly compared to the absolute magnitude of the c_{jkl} and the c_{kl} effects in Equation 4.3 to ascertain whether there is a greater potential for bias in the Bookmark or Angoff standard setting procedures. In fact, it is possible to use the item difficulty gaps from Equation 4.10 in a fixed effects model in the same way that the residuals and absolute residuals in Equation 4.3 are used to estimate the indices in Equations 4.10 and 4.13. That is, one can formulate the model:

$$\theta_{L_{jkl}} - \theta_{S_{jkl}} = \mathbf{x}_{jkl}\boldsymbol{\beta} + c_j + c_k + c_l + c_{jk} + c_{jl} + c_{kl} + c_{jkl} + u_{jkl}, \quad (4.16)$$

and estimate the fixed effects c_{jkl} and the c_{kl} . These effects would be exactly the same as the indices in Equations 4.10 and 4.13. An important observation in applying the model in Equation 4.16 is that when the fixed effect for an individual panelist is estimated, the c_{jkl} effect, it is not possible to test the statistical significance of this effect since there is only one item difficulty gap at each performance level. This means that the random error term would be equal to zero. In such cases, the standard error is indeterminate.

In the PAC metric, the value of the indices in Equations 4.14 and 4.15 can be compared and contrasted with the values of Equation 4.6 and 4.7 to determine whether the Bookmark or Angoff procedure has a greater potential to change the number of students classified in each performance category.

In Chapter 5, data from the 2005 12th grade Mathematics NAEP pilot study is used to compute the Bookmark and Angoff indices in both the score scale metric and PAC metric for the Mapmark and Angoff procedure with Mean Estimation that were applied to set cut scores on these data. For the Mapmark procedure, which is essentially the Bookmark procedure in round 1, the value of these indices will be calculated for the Bookmark placements at each performance level in round 1 of the mathematics pilot study that used a RP value of 0.67. It is expected that the value of the indices will be quite small since the number of items used in the standard setting procedure was quite large and the items were fairly spread out across the score scale. The indices for the Mapmark procedure are then compared to indices for the Angoff procedure with Mean Estimation in round 1 using the same data. The full model in Equation 4.3 is also estimated with regular and absolute residuals for the Angoff procedure across all the rounds of standard setting to illustrate how the statistical models can be used over rounds. These models will help to provide some indication of how panelist inconsistency is impacted by feedback and panelist interactions over rounds.

The expectation is that the biases will be larger for the Angoff procedure with Mean Estimation than for the Mapmark procedure since it is believed that the cognitive complexity of the Angoff procedure is a bigger problem than the item difficulty gaps in Bookmark. It is also expected that the potential biases in the 2005 mathematics NAEP

pilot study will be quite small for each method and that the biases will be smaller in later rounds of the Angoff procedure since panelists have more experience and training using the standard setting methods in later rounds of the process.

CHAPTER 5

COMPARISON OF ANGOFF AND MAPMARK METHODS

In this chapter, the indices that were developed in Chapter 4 are applied to data from the 2005 12th grade mathematics pilot study of NAEP in which a version of the Angoff and Mapmark (a version of the Bookmark procedure with different feedback in rounds 2, 3, and 4) procedures were applied with two nationally representative and equivalent groups of panelists to set cut scores. First, the datasets and standard setting procedures analyzed in this chapter are described in further detail. The indices in Chapter 4 are then calculated in scale score metric and PAC metric and the results from the two standard setting procedures are compared. The chapter concludes with a discussion of the empirical comparison findings.

5.1 Description of 2005 NAEP Mathematics Pilot Study Data and Procedures

The data set used is from a pilot study of the NAEP Grade 12 mathematics standard setting. In this pilot study, two different standard setting procedures were tried out with two nationally representative samples of panelists. The items used in the standard setting consisted of multiple-choice, short-answer, and constructed-response items that were field tested in 2004 and were later included in the NAEP item pool. Each of these items was fit with the 2PL, 3PL, or GPCM.

One of the standard setting methods was an Angoff item rating method with Mean Estimation. This standard setting procedure had been used previously in other NAEP standard settings and consists of four rounds of ratings. In the first round, panelists

discussed what students should know and be able to do at each performance level and provide their initial Angoff ratings. In the second round, panelists received feedback in the form of a Reckase chart (Reckase, 2001), conditional p -values at their cut scores, and the location of their cut score in relationship to other panelists. Panelists discussed this information and gave a second round of ratings. In the third round, panelists received all of the same feedback as round 2 and additional feedback based on how many students would be above their own cut scores. After reviewing this information, they provided a third round of ratings. In the fourth round, panelists received information on the percent above the cut score for the cut score of the whole group. They then indicated a cut score on the score scale used in the standard setting as their fourth round of ratings.

Data from the Angoff standard setting included information on the standard setting judgments of the twenty panelists that participated in each round, the location of their cut scores and their ratings, and the item parameters for each of the items. Throughout the Angoff standard setting process, the twenty panelists were divided into two independent groups of ten panelists who each performed the Angoff standard setting process separately. These two groups of ten panelists were further subdivided into two table groups within each replication. These table groups allowed the panelists to discuss their ratings and receive feedback from other panelists in between rounds. Panelists were instructed to independently indicate their ratings on each of the items. The first group of panelists rated 107 NAEP items and the second group of panelists rated 109 NAEP items with 39 common items across the two groups.

The second standard setting procedure used in the pilot study was the Mapmark method. In the first round, the Mapmark method is essentially the same as the Bookmark

procedure since panelists are asked to go through a set of ordered items with a RP value of 0.67. They have to place a bookmark between the items that separate what students in each of the different performance levels should know and be able to do from the items that students should not know and be unable to do with a greater than 67 percent probability. In the second and later rounds, the Mapmark method diverges from the Bookmark method in that panelists are given feedback on the expected performance of students at their cut score estimates across a set of teacher domains in a domain score chart and panelists are allowed to use this information to set their cut scores. Rater location data and condensed item maps are also used in later rounds. Hence, the Mapmark differs from the Bookmark method in that panelists do not indicate their cut score by placing a mark in an OIB in these later rounds, but rather they indicate their cut score on a domain score chart and answer questions of whether they think the cut score should be higher or lower based on this information.

This change in the standard setting task in the later rounds of Mapmark compared to the traditional Bookmark method means that different methods would be needed to evaluate the procedure in the first round compared to the later rounds of standard setting. Therefore, the comparison used in this study will focus on the first round of the Mapmark procedure where the standard setting task is essentially the same as the Bookmark method. In addition, the differences between the ratings in the first round of the Angoff method and the Mapmark method and how Reckase's initial psychometric theory could be applied in this situation was a point of major contention raised by Schulz (2006).

Similar to the Angoff standard setting process, information is available on the twenty-one panelists that participated in each round of the Mapmark procedure, the

location of their cut scores, and the item parameters for each of the items. Throughout the Mapmark standard setting process, the twenty-one panelists were again divided into two independent groups, one group of ten panelists and one group of eleven panelists. These two groups of ten and eleven panelists were further subdivided into two table groups within each replication. Each panelist was again instructed to provide independent standard setting judgments in each round. The first group of panelists again rated 107 NAEP items which were presented in an OIB and the second group of panelists rated 109 NAEP items in an OIB. Again, 39 of the items rated overlapped. Further information on these two sets of data can be found in 2005 mathematics standard setting special studies report (ACT, 2005).

5.2 Analysis Procedures

In the analyses that follow, the Angoff ratings in the first three rounds are used to estimate the models and indices presented in Chapter 4. The fourth round of ratings is excluded from the analyses since the panelists did not actually provide Angoff ratings in the fourth round and instead just indicated where they thought the cut score should be located. For the Mapmark method only the first round of bookmark placements will be analyzed. The results from the first round of the Mapmark are compared with the Angoff ratings in the first round. Throughout the analyses, the scale score values that correspond to item ratings in the Angoff procedure with Mean Estimation or the bookmark placement in the Mapmark procedure will be used in the computation of the indices from Chapter 4. These scale score values are simple linear transformations of IRT θ estimates.

Thus, the score values are used in place of θ in the analyses. The score scale used in the mathematics pilot study ranged from 100 to 400.

To determine the value of the potential individual panelist biases and inconsistencies as well as the potential biases and inconsistencies for the group of panelists for the Angoff method, the fixed effects model in Equation 4.3 is fit to these data with the residuals and absolute residuals as dependent variables and no covariates except for the c_{jkl} or c_{kl} fixed effects, respectively. The residuals are calculated by subtracting the score scale value for item rating from the estimated score scale value for the estimated cut score of that panelist. That is, if the panelist's item rating for an item corresponds to a score scale of 286 and the estimated cut score for that panelist is 291, then the estimated residual for that panelist for that item would be -5. The absolute residuals are found by taking the absolute value of the residuals. In this case, the absolute residual would be 5.

The c_{jkl} or c_{kl} fixed effects are estimated using the least-squares dummy-variable regression model without an intercept. This allows each of the coefficients on the fixed effects to be interpreted as the average potential bias or inconsistency for the effect of interest. Bias in this sense is defined as the average difference between the item ratings and the estimated cut score.

For example, if the effect is c_{111} with the residuals as dependent variables this effect would be interpreted as the potential bias for the first panelist in the first round for the first cut score estimate (e.g., the basic level). If the effect is positive, this indicates that the item ratings are too high compared to the cut score estimated from the panelist's ratings. Conversely, if the effect is negative, this indicates that the item ratings are too

low compared to the panelist's estimated cut score. With the absolute residuals, the effect c_{111} would be interpreted as the average absolute inconsistency for the first panelist in the first round for the first cut score estimate. These effects can be directly tested to see if the potential bias or inconsistency for this panelist in this round at this performance level is or is not statistically significant different from zero. The estimated effects from the model with residuals as dependent variables are then added to the estimated cut score for that panelist or group of panelists and the PAC for the original cut score estimate and the estimate that includes the potential bias from the fixed effects models are calculated and compared.

A similar approach is taken to estimate the indices for the Mapmark procedure represented in Equations 4.10 and 4.13 as is used with the Angoff procedure. In this case, the dependent variables in the least-squares dummy-variable regression model with no intercept are the item difficulty gaps. The item difficulty gaps are calculated from subtracting the scale score value for the item that the panelist bookmarked in the OIB from the next highest scale score value in the OIB. These item difficulty gaps are always positive, which means that there is the potential for the cut score estimates in the Mapmark method to be underestimated compared to the intended cut score. This is in contrast to the Angoff method, where it is possible for the cut score estimate to be perfect or over-or underestimated depending on the ratings given by the panelists.

The fixed effects of interest in the model again are the c_{jkl} or c_{kl} effects. For the c_{jkl} effects, each panelist only has one item difficulty gap at each performance level in each round, which means that it is not possible to statistically test these effects to see if they are statistically different from zero since the standard error is indeterminate. The c_{kl}

effects can be statistically tested to see if they are different from zero since these effects are pooled over panelists within a round and there is more than a single panelist that participates in the standard setting. The interpretation of the effects from these models are similar to the interpretation of the effects with the Angoff procedure, where the c_{111} effect, for example, would be interpreted as the maximum potential bias for the first panelist in the first round for the first cut score estimate (e.g., the basic level). This value is simply the item difficulty gap in Equation 4.10. If the effect was c_{11} this would be the average maximum potential bias for the first cut score estimate in the first round for the group of panelists. Again, the estimated effects from the fixed effects models are added to the estimated cut score for that panelist or the group of panelists and the PAC for the original cut score estimate and the estimate that includes the potential bias are calculated and compared.

5.3 Results for the Angoff Method

The results from fitting the fixed effects model with the residuals as dependent variables and the c_{jkl} effects are presented in Table 5.1. The estimates of the effects suggest that there is the potential for bias in the individual panelist judgments since the effects for each of the panelists at each cut score in each round are not universally equal to zero. In some cases, the estimated effects are negative, indicating that the cut scores are overestimated compared with what the ratings should be if the panelist gave ratings completely in line with their estimated cut score. In other situations, the effects are positive, indicating that the cut scores are underestimated compared to what the ratings should be if the panelist gave ratings completely inline with their estimated cut score. For

example, the estimated effect for panelist A1204 in round 1 at the basic level is -12.811 meaning that the average item rating provided by this panelist is 12.811 scale score points lower than their estimated cut score at this performance level in this round.

The effects for the different panelists at each performance level in each round are varied. This variation indicates that the potential biases can change depending on the panelist and the situation in which the panelists are asked to provide their ratings. In general, the magnitude of the biases for the panelists tends to be greatest for the basic level and in earlier rounds of the standard setting process, although there are some other potentially large biases for other levels and rounds of the process for particular panelists.

Table 5.1: Potential Panelist Biases in Angoff Ratings

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
A1201	R1	Basic	-2.887	2.998	-0.963	0.336
A1202	R1	Basic	-11.000	2.998	-3.669	0.000
A1203	R1	Basic	-9.094	2.998	-3.033	0.002
A1204	R1	Basic	-12.821	2.998	-4.276	0.000
A1205	R1	Basic	2.906	2.998	0.969	0.333
A1206	R1	Basic	-5.123	2.998	-1.709	0.088
A1207	R1	Basic	-20.406	2.998	-6.806	0.000
A1208	R1	Basic	-17.830	2.998	-5.947	0.000
A1209	R1	Basic	-19.651	2.998	-6.554	0.000
A1210	R1	Basic	0.972	2.998	0.324	0.746
B1211	R1	Basic	-8.454	2.970	-2.846	0.004
B1212	R1	Basic	-15.185	2.970	-5.112	0.000
B1213	R1	Basic	-5.861	2.970	-1.973	0.048
B1214	R1	Basic	-4.546	2.970	-1.531	0.126
B1215	R1	Basic	-7.259	2.970	-2.444	0.015
B1216	R1	Basic	-8.574	2.970	-2.886	0.004
B1217	R1	Basic	-6.491	2.970	-2.185	0.029
B1218	R1	Basic	-3.732	2.970	-1.256	0.209
B1219	R1	Basic	-19.361	2.970	-6.518	0.000
B1220	R1	Basic	-3.019	2.970	-1.016	0.310
A1201	R2	Basic	-5.208	2.998	-1.737	0.082
A1202	R2	Basic	-4.962	2.998	-1.655	0.098
A1203	R2	Basic	-7.283	2.998	-2.429	0.015
A1204	R2	Basic	-8.425	2.998	-2.810	0.005
A1205	R2	Basic	2.076	2.998	0.692	0.489
A1206	R2	Basic	-0.736	2.998	-0.245	0.806
A1207	R2	Basic	-4.623	2.998	-1.542	0.123

Table 5.1 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
A1208	R2	Basic	-1.774	2.998	-0.592	0.554
A1209	R2	Basic	-10.840	2.998	-3.615	0.000
A1210	R2	Basic	2.689	2.998	0.897	0.370
B1211	R2	Basic	-0.630	2.970	-0.212	0.832
B1212	R2	Basic	0.176	2.970	0.059	0.953
B1213	R2	Basic	-3.157	2.970	-1.063	0.288
B1214	R2	Basic	0.472	2.970	0.159	0.874
B1215	R2	Basic	-28.982	2.970	-9.757	0.000
B1216	R2	Basic	-0.380	2.970	-0.128	0.898
B1217	R2	Basic	-0.963	2.970	-0.324	0.746
B1218	R2	Basic	1.565	2.970	0.527	0.598
B1219	R2	Basic	-10.750	2.970	-3.619	0.000
B1220	R2	Basic	-2.269	2.970	-0.764	0.445
A1201	R3	Basic	-5.566	2.998	-1.856	0.063
A1202	R3	Basic	-5.576	2.998	-1.860	0.063
A1203	R3	Basic	-14.991	2.998	-5.000	0.000
A1204	R3	Basic	-6.547	2.998	-2.184	0.029
A1205	R3	Basic	2.047	2.998	0.683	0.495
A1206	R3	Basic	1.868	2.998	0.623	0.533
A1207	R3	Basic	-1.311	2.998	-0.437	0.662
A1208	R3	Basic	-4.028	2.998	-1.344	0.179
A1209	R3	Basic	-2.189	2.998	-0.730	0.465
A1210	R3	Basic	2.321	2.998	0.774	0.439
B1211	R3	Basic	0.324	2.970	0.109	0.913
B1212	R3	Basic	0.676	2.970	0.228	0.820
B1213	R3	Basic	-3.407	2.970	-1.147	0.251
B1214	R3	Basic	-0.472	2.970	-0.159	0.874
B1215	R3	Basic	-25.639	2.970	-8.631	0.000
B1216	R3	Basic	-7.482	2.970	-2.519	0.012
B1217	R3	Basic	-6.843	2.970	-2.304	0.021
B1218	R3	Basic	-6.380	2.970	-2.148	0.032
B1219	R3	Basic	-7.278	2.970	-2.450	0.014
B1220	R3	Basic	-3.093	2.970	-1.041	0.298
A1201	R1	Proficient	-1.123	2.998	-0.374	0.708
A1202	R1	Proficient	6.519	2.998	2.174	0.030
A1203	R1	Proficient	0.321	2.998	0.107	0.915
A1204	R1	Proficient	-6.425	2.998	-2.143	0.032
A1205	R1	Proficient	10.038	2.998	3.348	0.001
A1206	R1	Proficient	2.406	2.998	0.802	0.422
A1207	R1	Proficient	2.132	2.998	0.711	0.477
A1208	R1	Proficient	-4.500	2.998	-1.501	0.133
A1209	R1	Proficient	-0.566	2.998	-0.189	0.850
A1210	R1	Proficient	10.085	2.998	3.364	0.001
B1211	R1	Proficient	1.611	2.970	0.542	0.588
B1212	R1	Proficient	-3.926	2.970	-1.322	0.186

Table 5.1 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
B1213	R1	Proficient	-4.056	2.970	-1.365	0.172
B1214	R1	Proficient	-4.037	2.970	-1.359	0.174
B1215	R1	Proficient	-4.806	2.970	-1.618	0.106
B1216	R1	Proficient	2.870	2.970	0.966	0.334
B1217	R1	Proficient	-7.972	2.970	-2.684	0.007
B1218	R1	Proficient	4.528	2.970	1.524	0.127
B1219	R1	Proficient	-2.880	2.970	-0.969	0.332
B1220	R1	Proficient	-4.537	2.970	-1.527	0.127
A1201	R2	Proficient	-3.887	2.998	-1.296	0.195
A1202	R2	Proficient	5.066	2.998	1.690	0.091
A1203	R2	Proficient	0.094	2.998	0.031	0.975
A1204	R2	Proficient	-0.943	2.998	-0.315	0.753
A1205	R2	Proficient	7.236	2.998	2.413	0.016
A1206	R2	Proficient	3.151	2.998	1.051	0.293
A1207	R2	Proficient	1.745	2.998	0.582	0.561
A1208	R2	Proficient	0.076	2.998	0.025	0.980
A1209	R2	Proficient	0.293	2.998	0.098	0.922
A1210	R2	Proficient	8.491	2.998	2.832	0.005
B1211	R2	Proficient	1.287	2.970	0.433	0.665
B1212	R2	Proficient	-2.167	2.970	-0.729	0.466
B1213	R2	Proficient	-4.602	2.970	-1.549	0.121
B1214	R2	Proficient	-2.815	2.970	-0.948	0.343
B1215	R2	Proficient	-5.769	2.970	-1.942	0.052
B1216	R2	Proficient	-2.019	2.970	-0.680	0.497
B1217	R2	Proficient	-3.815	2.970	-1.284	0.199
B1218	R2	Proficient	-0.546	2.970	-0.184	0.854
B1219	R2	Proficient	-2.482	2.970	-0.835	0.404
B1220	R2	Proficient	-5.611	2.970	-1.889	0.059
A1201	R3	Proficient	-3.906	2.998	-1.303	0.193
A1202	R3	Proficient	4.255	2.998	1.419	0.156
A1203	R3	Proficient	-2.113	2.998	-0.705	0.481
A1204	R3	Proficient	-0.783	2.998	-0.261	0.794
A1205	R3	Proficient	5.981	2.998	1.995	0.046
A1206	R3	Proficient	2.293	2.998	0.765	0.445
A1207	R3	Proficient	0.887	2.998	0.296	0.767
A1208	R3	Proficient	0.349	2.998	0.116	0.907
A1209	R3	Proficient	0.717	2.998	0.239	0.811
A1210	R3	Proficient	6.585	2.998	2.196	0.028
B1211	R3	Proficient	2.241	2.970	0.754	0.451
B1212	R3	Proficient	-2.917	2.970	-0.982	0.326
B1213	R3	Proficient	-2.796	2.970	-0.941	0.347
B1214	R3	Proficient	-1.324	2.970	-0.446	0.656
B1215	R3	Proficient	-3.241	2.970	-1.091	0.275
B1216	R3	Proficient	-0.278	2.970	-0.094	0.925
B1217	R3	Proficient	-4.843	2.970	-1.630	0.103

Table 5.1 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
B1218	R3	Proficient	1.019	2.970	0.343	0.732
B1219	R3	Proficient	-2.093	2.970	-0.704	0.481
B1220	R3	Proficient	-6.870	2.970	-2.313	0.021
A1201	R1	Advanced	-1.151	2.998	-0.384	0.701
A1202	R1	Advanced	5.717	2.998	1.907	0.057
A1203	R1	Advanced	-6.340	2.998	-2.114	0.034
A1204	R1	Advanced	-3.179	2.998	-1.060	0.289
A1205	R1	Advanced	18.123	2.998	6.044	0.000
A1206	R1	Advanced	9.085	2.998	3.030	0.002
A1207	R1	Advanced	3.057	2.998	1.019	0.308
A1208	R1	Advanced	-1.670	2.998	-0.557	0.578
A1209	R1	Advanced	9.623	2.998	3.209	0.001
A1210	R1	Advanced	9.613	2.998	3.206	0.001
B1211	R1	Advanced	3.167	2.970	1.066	0.286
B1212	R1	Advanced	1.537	2.970	0.517	0.605
B1213	R1	Advanced	-1.620	2.970	-0.546	0.585
B1214	R1	Advanced	-0.630	2.970	-0.212	0.832
B1215	R1	Advanced	-4.593	2.970	-1.546	0.122
B1216	R1	Advanced	-4.370	2.970	-1.471	0.141
B1217	R1	Advanced	-10.111	2.970	-3.404	0.001
B1218	R1	Advanced	-0.426	2.970	-0.143	0.886
B1219	R1	Advanced	-2.556	2.970	-0.860	0.390
B1220	R1	Advanced	-4.657	2.970	-1.568	0.117
A1201	R2	Advanced	-2.868	2.998	-0.957	0.339
A1202	R2	Advanced	10.717	2.998	3.574	0.000
A1203	R2	Advanced	0.094	2.998	0.031	0.975
A1204	R2	Advanced	10.000	2.998	3.335	0.001
A1205	R2	Advanced	18.425	2.998	6.145	0.000
A1206	R2	Advanced	5.726	2.998	1.910	0.056
A1207	R2	Advanced	5.226	2.998	1.743	0.081
A1208	R2	Advanced	1.208	2.998	0.403	0.687
A1209	R2	Advanced	2.962	2.998	0.988	0.323
A1210	R2	Advanced	7.736	2.998	2.580	0.010
B1211	R2	Advanced	2.417	2.970	0.814	0.416
B1212	R2	Advanced	0.991	2.970	0.334	0.739
B1213	R2	Advanced	-0.213	2.970	-0.072	0.943
B1214	R2	Advanced	-1.769	2.970	-0.595	0.552
B1215	R2	Advanced	2.074	2.970	0.698	0.485
B1216	R2	Advanced	-4.972	2.970	-1.674	0.094
B1217	R2	Advanced	-4.482	2.970	-1.509	0.131
B1218	R2	Advanced	-1.398	2.970	-0.471	0.638
B1219	R2	Advanced	-0.750	2.970	-0.252	0.801
B1220	R2	Advanced	-5.139	2.970	-1.730	0.084
A1201	R3	Advanced	-2.896	2.998	-0.966	0.334

Table 5.1 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
A1202	R3	Advanced	12.566	2.998	4.191	0.000
A1203	R3	Advanced	-0.057	2.998	-0.019	0.985
A1204	R3	Advanced	11.179	2.998	3.729	0.000
A1205	R3	Advanced	14.462	2.998	4.823	0.000
A1206	R3	Advanced	3.406	2.998	1.136	0.256
A1207	R3	Advanced	5.557	2.998	1.853	0.064
A1208	R3	Advanced	1.859	2.998	0.620	0.535
A1209	R3	Advanced	3.953	2.998	1.318	0.187
A1210	R3	Advanced	4.972	2.998	1.658	0.097
B1211	R3	Advanced	2.732	2.970	0.920	0.358
B1212	R3	Advanced	0.296	2.970	0.100	0.921
B1213	R3	Advanced	9.454	2.970	3.183	0.001
B1214	R3	Advanced	-2.259	2.970	-0.761	0.447
B1215	R3	Advanced	3.157	2.970	1.063	0.288
B1216	R3	Advanced	-2.602	2.970	-0.876	0.381
B1217	R3	Advanced	-6.065	2.970	-2.042	0.041
B1218	R3	Advanced	-2.398	2.970	-0.807	0.419
B1219	R3	Advanced	-0.750	2.970	-0.252	0.801
B1220	R3	Advanced	-6.241	2.970	-2.101	0.036

Complimenting the investigations of the potential biases for the individual panelists are the investigations of the potential inconsistencies of the same panelists in the same situations, which are obtained by applying the fixed models with the absolute residuals. The results of this analysis are displayed in Table 5.2. Again, the estimates of the effects are not universally equal to zero indicating that panelists are not completely consistent in their standard setting judgments. For example, panelist A1201 has an estimated inconsistency of 33.755 for the basic level in round 1. This level of inconsistency suggests that the average absolute deviation of the individual item ratings for this panelist from their estimated cut score estimate for this performance level in this round is 33.755 scale score points.

An important pattern observed is that the panelist inconsistencies tend to be larger in earlier rounds of the Angoff standard setting process and less in the later rounds of the

process (Table 5.2). Each of these inconsistencies is statistically significantly different from zero. Similar to the findings with the potential biases for the individual panelists, the inconsistencies for the panelists tend to be largest at the basic level in comparison to the other performance levels. Combining Table 5.1 and Table 5.2 together suggests that not only are panelists inconsistent in their ratings, but that there is also the potential for biases in the cut score estimates for the individual panelists from the panelist inconsistencies.

Table 5.2: Panelist Inconsistencies in Angoff Ratings

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
A1201	R1	Basic	33.755	2.180	15.487	0.000
A1202	R1	Basic	38.151	2.180	17.504	0.000
A1203	R1	Basic	33.585	2.180	15.409	0.000
A1204	R1	Basic	37.915	2.180	17.396	0.000
A1205	R1	Basic	29.132	2.180	13.366	0.000
A1206	R1	Basic	30.821	2.180	14.141	0.000
A1207	R1	Basic	45.802	2.180	21.015	0.000
A1208	R1	Basic	39.132	2.180	17.954	0.000
A1209	R1	Basic	43.802	2.180	20.097	0.000
A1210	R1	Basic	26.274	2.180	12.055	0.000
B1211	R1	Basic	41.139	2.159	19.052	0.000
B1212	R1	Basic	56.056	2.159	25.961	0.000
B1213	R1	Basic	32.731	2.159	15.159	0.000
B1214	R1	Basic	28.361	2.159	13.135	0.000
B1215	R1	Basic	35.185	2.159	16.295	0.000
B1216	R1	Basic	38.093	2.159	17.642	0.000
B1217	R1	Basic	31.546	2.159	14.610	0.000
B1218	R1	Basic	32.694	2.159	15.142	0.000
B1219	R1	Basic	45.694	2.159	21.162	0.000
B1220	R1	Basic	27.481	2.159	12.727	0.000
A1201	R2	Basic	27.057	2.180	12.414	0.000
A1202	R2	Basic	22.623	2.180	10.380	0.000
A1203	R2	Basic	20.038	2.180	9.194	0.000
A1204	R2	Basic	35.123	2.180	16.115	0.000
A1205	R2	Basic	27.509	2.180	12.622	0.000
A1206	R2	Basic	19.321	2.180	8.865	0.000
A1207	R2	Basic	8.094	2.180	3.714	0.000
A1208	R2	Basic	13.075	2.180	5.999	0.000
A1209	R2	Basic	18.708	2.180	8.583	0.000
A1210	R2	Basic	25.632	2.180	11.760	0.000
B1211	R2	Basic	21.630	2.159	10.017	0.000
B1212	R2	Basic	14.713	2.159	6.814	0.000
B1213	R2	Basic	25.398	2.159	11.762	0.000

Table 5.2 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
B1214	R2	Basic	11.472	2.159	5.313	0.000
B1215	R2	Basic	44.574	2.159	20.643	0.000
B1216	R2	Basic	9.880	2.159	4.575	0.000
B1217	R2	Basic	21.315	2.159	9.871	0.000
B1218	R2	Basic	18.917	2.159	8.761	0.000
B1219	R2	Basic	33.343	2.159	15.442	0.000
B1220	R2	Basic	22.157	2.159	10.262	0.000
A1201	R3	Basic	27.245	2.180	12.501	0.000
A1202	R3	Basic	19.368	2.180	8.886	0.000
A1203	R3	Basic	24.500	2.180	11.241	0.000
A1204	R3	Basic	32.321	2.180	14.829	0.000
A1205	R3	Basic	19.915	2.180	9.137	0.000
A1206	R3	Basic	13.566	2.180	6.224	0.000
A1207	R3	Basic	4.142	2.180	1.900	0.057
A1208	R3	Basic	11.443	2.180	5.250	0.000
A1209	R3	Basic	6.170	2.180	2.831	0.005
A1210	R3	Basic	24.094	2.180	11.055	0.000
B1211	R3	Basic	14.083	2.159	6.522	0.000
B1212	R3	Basic	11.509	2.159	5.330	0.000
B1213	R3	Basic	23.907	2.159	11.072	0.000
B1214	R3	Basic	11.546	2.159	5.347	0.000
B1215	R3	Basic	35.991	2.159	16.668	0.000
B1216	R3	Basic	14.685	2.159	6.801	0.000
B1217	R3	Basic	22.361	2.159	10.356	0.000
B1218	R3	Basic	13.546	2.159	6.274	0.000
B1219	R3	Basic	29.944	2.159	13.868	0.000
B1220	R3	Basic	20.352	2.159	9.425	0.000
A1201	R1	Proficient	23.330	2.180	10.704	0.000
A1202	R1	Proficient	20.953	2.180	9.613	0.000
A1203	R1	Proficient	21.887	2.180	10.042	0.000
A1204	R1	Proficient	22.311	2.180	10.237	0.000
A1205	R1	Proficient	26.792	2.180	12.293	0.000
A1206	R1	Proficient	21.066	2.180	9.665	0.000
A1207	R1	Proficient	25.132	2.180	11.531	0.000
A1208	R1	Proficient	21.764	2.180	9.986	0.000
A1209	R1	Proficient	23.208	2.180	10.648	0.000
A1210	R1	Proficient	25.349	2.180	11.631	0.000
B1211	R1	Proficient	27.463	2.159	12.719	0.000
B1212	R1	Proficient	25.944	2.159	12.015	0.000
B1213	R1	Proficient	20.704	2.159	9.588	0.000
B1214	R1	Proficient	22.093	2.159	10.232	0.000
B1215	R1	Proficient	21.731	2.159	10.064	0.000
B1216	R1	Proficient	25.056	2.159	11.604	0.000
B1217	R1	Proficient	22.991	2.159	10.648	0.000

Table 5.2 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
B1218	R1	Proficient	23.324	2.159	10.802	0.000
B1219	R1	Proficient	23.398	2.159	10.836	0.000
B1220	R1	Proficient	23.426	2.159	10.849	0.000
A1201	R2	Proficient	18.302	2.180	8.397	0.000
A1202	R2	Proficient	13.651	2.180	6.263	0.000
A1203	R2	Proficient	12.415	2.180	5.696	0.000
A1204	R2	Proficient	20.321	2.180	9.323	0.000
A1205	R2	Proficient	21.915	2.180	10.055	0.000
A1206	R2	Proficient	15.925	2.180	7.306	0.000
A1207	R2	Proficient	5.745	2.180	2.636	0.008
A1208	R2	Proficient	7.396	2.180	3.394	0.001
A1209	R2	Proficient	8.368	2.180	3.839	0.000
A1210	R2	Proficient	24.377	2.180	11.185	0.000
B1211	R2	Proficient	16.454	2.159	7.620	0.000
B1212	R2	Proficient	12.074	2.159	5.592	0.000
B1213	R2	Proficient	17.954	2.159	8.315	0.000
B1214	R2	Proficient	9.333	2.159	4.322	0.000
B1215	R2	Proficient	16.880	2.159	7.817	0.000
B1216	R2	Proficient	5.796	2.159	2.684	0.007
B1217	R2	Proficient	17.574	2.159	8.139	0.000
B1218	R2	Proficient	13.454	2.159	6.231	0.000
B1219	R2	Proficient	18.926	2.159	8.765	0.000
B1220	R2	Proficient	19.574	2.159	9.065	0.000
A1201	R3	Proficient	18.283	2.180	8.389	0.000
A1202	R3	Proficient	12.179	2.180	5.588	0.000
A1203	R3	Proficient	11.170	2.180	5.125	0.000
A1204	R3	Proficient	20.274	2.180	9.302	0.000
A1205	R3	Proficient	17.321	2.180	7.947	0.000
A1206	R3	Proficient	14.047	2.180	6.445	0.000
A1207	R3	Proficient	2.925	2.180	1.342	0.180
A1208	R3	Proficient	5.500	2.180	2.523	0.012
A1209	R3	Proficient	2.868	2.180	1.316	0.188
A1210	R3	Proficient	22.943	2.180	10.527	0.000
B1211	R3	Proficient	11.241	2.159	5.206	0.000
B1212	R3	Proficient	10.102	2.159	4.678	0.000
B1213	R3	Proficient	17.259	2.159	7.993	0.000
B1214	R3	Proficient	6.139	2.159	2.843	0.004
B1215	R3	Proficient	11.481	2.159	5.317	0.000
B1216	R3	Proficient	3.593	2.159	1.664	0.096
B1217	R3	Proficient	14.972	2.159	6.934	0.000
B1218	R3	Proficient	4.667	2.159	2.161	0.031
B1219	R3	Proficient	18.389	2.159	8.516	0.000
B1220	R3	Proficient	18.537	2.159	8.585	0.000
A1201	R1	Advanced	19.358	2.180	8.882	0.000
A1202	R1	Advanced	21.755	2.180	9.981	0.000

Table 5.2 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
A1203	R1	Advanced	22.358	2.180	10.258	0.000
A1204	R1	Advanced	22.481	2.180	10.315	0.000
A1205	R1	Advanced	28.613	2.180	13.128	0.000
A1206	R1	Advanced	21.840	2.180	10.020	0.000
A1207	R1	Advanced	27.849	2.180	12.778	0.000
A1208	R1	Advanced	22.142	2.180	10.159	0.000
A1209	R1	Advanced	24.755	2.180	11.358	0.000
A1210	R1	Advanced	24.953	2.180	11.449	0.000
B1211	R1	Advanced	26.500	2.159	12.273	0.000
B1212	R1	Advanced	26.074	2.159	12.076	0.000
B1213	R1	Advanced	21.528	2.159	9.970	0.000
B1214	R1	Advanced	19.481	2.159	9.022	0.000
B1215	R1	Advanced	23.093	2.159	10.695	0.000
B1216	R1	Advanced	24.037	2.159	11.132	0.000
B1217	R1	Advanced	26.037	2.159	12.058	0.000
B1218	R1	Advanced	24.130	2.159	11.175	0.000
B1219	R1	Advanced	23.167	2.159	10.729	0.000
B1220	R1	Advanced	22.954	2.159	10.630	0.000
A1201	R2	Advanced	16.585	2.180	7.609	0.000
A1202	R2	Advanced	15.189	2.180	6.969	0.000
A1203	R2	Advanced	16.528	2.180	7.583	0.000
A1204	R2	Advanced	20.132	2.180	9.237	0.000
A1205	R2	Advanced	27.538	2.180	12.635	0.000
A1206	R2	Advanced	17.028	2.180	7.813	0.000
A1207	R2	Advanced	11.453	2.180	5.255	0.000
A1208	R2	Advanced	7.943	2.180	3.645	0.000
A1209	R2	Advanced	7.811	2.180	3.584	0.000
A1210	R2	Advanced	21.604	2.180	9.912	0.000
B1211	R2	Advanced	16.694	2.159	7.732	0.000
B1212	R2	Advanced	12.009	2.159	5.562	0.000
B1213	R2	Advanced	18.824	2.159	8.718	0.000
B1214	R2	Advanced	11.028	2.159	5.107	0.000
B1215	R2	Advanced	19.130	2.159	8.859	0.000
B1216	R2	Advanced	8.991	2.159	4.164	0.000
B1217	R2	Advanced	21.296	2.159	9.863	0.000
B1218	R2	Advanced	13.880	2.159	6.428	0.000
B1219	R2	Advanced	22.213	2.159	10.287	0.000
B1220	R2	Advanced	21.120	2.159	9.781	0.000
A1201	R3	Advanced	16.557	2.180	7.596	0.000
A1202	R3	Advanced	14.868	2.180	6.822	0.000
A1203	R3	Advanced	9.453	2.180	4.337	0.000
A1204	R3	Advanced	20.009	2.180	9.181	0.000
A1205	R3	Advanced	21.745	2.180	9.977	0.000
A1206	R3	Advanced	14.651	2.180	6.722	0.000

Table 5.2 (cont'd)

Rater	Round	Cut	Estimate	S.E.	t-value	Pr(> t)
A1207	R3	Advanced	10.274	2.180	4.714	0.000
A1208	R3	Advanced	6.519	2.180	2.991	0.003
A1209	R3	Advanced	5.500	2.180	2.523	0.012
A1210	R3	Advanced	19.764	2.180	9.068	0.000
B1211	R3	Advanced	11.250	2.159	5.210	0.000
B1212	R3	Advanced	10.981	2.159	5.086	0.000
B1213	R3	Advanced	19.231	2.159	8.907	0.000
B1214	R3	Advanced	6.778	2.159	3.139	0.002
B1215	R3	Advanced	13.787	2.159	6.385	0.000
B1216	R3	Advanced	4.176	2.159	1.934	0.053
B1217	R3	Advanced	17.824	2.159	8.255	0.000
B1218	R3	Advanced	5.491	2.159	2.543	0.011
B1219	R3	Advanced	22.213	2.159	10.287	0.000
B1220	R3	Advanced	19.889	2.159	9.211	0.000

The results for the potential biases and inconsistencies across the group of panelists are displayed in Tables 5.3 and 5.4, respectively. The results in Tables 5.3 and 5.4 are just the aggregated effects for the panelist biases and inconsistencies presented in Tables 5.1 and 5.2.

Table 5.3: Potential Group Biases for Angoff Method

Round	Cut	Estimate	S.E.	t-value	Pr(> t)
R1	Basic	-8.865	0.676	-13.119	0.000
R2	Basic	-4.203	0.676	-6.220	0.000
R3	Basic	-4.690	0.676	-6.941	0.000
R1	Proficient	-0.236	0.676	-0.349	0.727
R2	Proficient	-0.384	0.676	-0.568	0.570
R3	Proficient	-0.358	0.676	-0.530	0.596
R1	Advanced	0.900	0.676	1.331	0.183
R2	Advanced	2.265	0.676	3.353	0.001
R3	Advanced	2.488	0.676	3.682	0.000

Table 5.4: Potential Group Inconsistencies for the Angoff Method

Round	Cut	Estimate	S.E.	t	Pr(> t)
R1	Basic	36.372	0.501	72.660	0.000
R2	Basic	22.032	0.501	44.010	0.000
R3	Basic	19.042	0.501	38.040	0.000
R1	Proficient	23.398	0.501	46.740	0.000
R2	Proficient	14.822	0.501	29.610	0.000
R3	Proficient	12.189	0.501	24.350	0.000
R1	Advanced	23.656	0.501	47.260	0.000
R2	Advanced	16.351	0.501	32.670	0.000
R3	Advanced	13.544	0.501	27.060	0.000

The greatest amount of potential bias in the group cut score estimates occurs in the first round at the basic level, where the estimated effect is -8.865 signifying that the group cut score estimate is about 8.865 points too high compared with item ratings provided by the panelists (Table 5.3). The effects at the basic level decrease in subsequent rounds, but still remain significant. At the proficient level, the effects are not statistical significant from zero indicating that potential biases in cut score estimates for this performance level are minimal. At the advanced level, the potential biases in the group cut score estimates increase across rounds ranging between 2 and 3 points lower than the item ratings provided by the panelists in the second and third rounds of the Angoff standard setting.

In terms of the inconsistencies for the groups of panelists, there is a general pattern of decreasing inconsistencies across the rounds of the standard setting process (Table 5.4). Each of the estimated effects is significant indicating that the group of panelists struggle to rate consistently with their estimated cut score (Table 5.4). The magnitude of the inconsistency is the greatest at the basic level with the proficient and advanced levels exhibiting similar levels of inconsistency. At the basic level, the estimated inconsistencies are roughly 36, 22, and 19 points across the three rounds of

standard setting which are approximately 13, 8, and 6 points higher than the levels of the average deviations at the proficient and advanced levels.

When the results from Tables 5.3 and 5.4 are again combined this suggests that the group of panelists have a hard time rating completely in line with their cut score estimate and this lack of consistency can translate into potential biases in the estimated group cut scores.

5.4 Results for the Mapmark Method

The findings from round 1 of Mapmark procedure are shown in Table 5.5. Somewhat differently from the findings for the Angoff procedure, the estimated potential effects for the individual panelists are uniformly positive ranging between 1 and 4 scale score points (Table 5.5). These uniformly positive effects signify that the cut score may be too low compared to the largest cut score that a panelist could be conceptualizing when applying the Bookmark procedure. The actual potential biases could be anywhere between zero and the estimated value. This suggests that there is the potential for the Mapmark cut scores to be underestimated compared to the panelist's intended cut score when applying the Bookmark procedure. For example, rater MA1205 has an estimated effect of 3 at the basic level meaning that there is the potential for this panelist's cut score estimate to be underestimated by as much as 3 scale score points (Table 5.5).

In addition, there does not appear to be a drastically higher potential underestimation problem at one performance level than at the other (Table 5.5). This occurs because the regions that the panelists choose to place their bookmarks were in

regions where the item difficulty gaps between cut score and the next highest cut score where roughly the same (i.e., similar gaps in the OIB).

Table 5.5: Maximum Potential Panelist Biases for the Mapmark Method

Rater	Round	Cut	Estimate
MA1201	R1	Basic	1.000
MA1202	R1	Basic	1.000
MA1203	R1	Basic	1.000
MA1204	R1	Basic	2.000
MA1205	R1	Basic	3.000
MA1206	R1	Basic	1.000
MA1207	R1	Basic	1.000
MA1208	R1	Basic	2.000
MA1209	R1	Basic	1.000
MA1210	R1	Basic	1.000
MB1211	R1	Basic	1.000
MB1212	R1	Basic	3.000
MB1213	R1	Basic	2.000
MB1214	R1	Basic	2.000
MB1215	R1	Basic	1.000
MB1216	R1	Basic	2.000
MB1217	R1	Basic	2.000
MB1218	R1	Basic	2.000
MB1219	R1	Basic	1.000
MB1220	R1	Basic	2.000
MB1221	R1	Basic	3.000
MA1201	R1	Proficient	1.000
MA1202	R1	Proficient	3.000
MA1203	R1	Proficient	1.000
MA1204	R1	Proficient	4.000
MA1205	R1	Proficient	4.000
MA1206	R1	Proficient	1.000
MA1207	R1	Proficient	1.000
MA1208	R1	Proficient	1.000
MA1209	R1	Proficient	3.000
MA1210	R1	Proficient	3.000
MB1211	R1	Proficient	1.000
MB1212	R1	Proficient	2.000
MB1213	R1	Proficient	1.000
MB1214	R1	Proficient	1.000
MB1215	R1	Proficient	2.000
MB1216	R1	Proficient	2.000
MB1217	R1	Proficient	2.000
MB1218	R1	Proficient	1.000
MB1219	R1	Proficient	1.000

Table 5.5 (cont'd)

Rater	Round	Cut	Estimate
MB1220	R1	Proficient	2.000
MB1221	R1	Proficient	2.000
MA1201	R1	Advanced	1.000
MA1202	R1	Advanced	1.000
MA1203	R1	Advanced	2.000
MA1204	R1	Advanced	2.000
MA1205	R1	Advanced	1.000
MA1206	R1	Advanced	2.000
MA1207	R1	Advanced	1.000
MA1208	R1	Advanced	3.000
MA1209	R1	Advanced	1.000
MA1210	R1	Advanced	1.000
MB1211	R1	Advanced	2.000
MB1212	R1	Advanced	1.000
MB1213	R1	Advanced	1.000
MB1214	R1	Advanced	1.000
MB1215	R1	Advanced	1.000
MB1216	R1	Advanced	1.000
MB1217	R1	Advanced	1.000
MB1218	R1	Advanced	1.000
MB1219	R1	Advanced	3.000
MB1220	R1	Advanced	1.000
MB1221	R1	Advanced	1.000

The findings of the potential maximum biases at the individual panelist level can again be aggregated to the group level to determine the maximum potential biases in the group cut score estimate from the existence of item difficulty gaps in the OIB. The results from this analysis are presented in Table 5.6.

Table 5.6: Maximum Potential Group Biases for the Mapmark Method

Round	Cut	Estimate	S.E.	t-value	Pr(> t)
R1	Basic	1.667	0.179	9.332	0.000
R1	Proficient	1.857	0.179	10.398	0.000
R1	Advanced	1.381	0.179	7.732	0.000

The table shows the average potential maximum biases in the group cut score estimates at the three performance levels. Each of the estimated effects is statistically

significant and between 1 and 2 scale score points on the NAEP scale score. The estimated effects indicate that there is the potential for the group cut score estimates in each performance category to be uniformly underestimated by between 1 to 2 scale score points on the NAEP scale score.

5.5 Comparison of Potential Biases between Angoff and Mapmark Methods

The potential biases reported in Tables 5.1 and 5.3 can be directly compared to the maximum potential biases in the Mapmark procedure depicted in Tables 5.5 and 5.6. Results indicate that the potential biases when applying the Angoff procedure for these NAEP standard setting data can be quite a bit larger than the maximum potential bias at the individual panelist level when applying the Mapmark procedure. However, this is not true at every cut score placement since the average biases of some of the panelists in the first round of Angoff standard setting are estimated to be less than one scale score point and are not statistically significant and different than zero.

Another essential observation in comparing the individual potential biases of panelists in the two procedures is that it is possible for the estimated potential biases in the Angoff method to be zero, negative, or positive, whereas in the Mapmark method the maximum potential biases are always positive due to the way in which the cut score is estimated. This suggests that if panelists perfectly understood the Angoff method and were able to carry it out accurately that the biases from applying the Angoff method would be less than the bias for the Mapmark procedure. This occurs since it is possible for the biases to turn out to be zero with the Angoff procedure, but it is not possible for

them to turn out to be zero with the Mapmark procedure since there will always be item difficulty gaps when applying the traditional Bookmark procedure.

At the group level, the biases for the Angoff procedure are somewhat larger in absolute magnitude for the basic level and somewhat smaller in absolute magnitude at the proficient and advanced levels than the Mapmark method. The smaller level of biases in the Angoff procedure at the proficient and advanced levels occurs despite the fact that several of the absolute magnitudes of the biases of the individual panelists are larger than the absolute magnitude of the biases of the panelists in the Mapmark procedure because the average of negative and positive panelist biases can cancel out and mitigate the amount of bias in the group cut score estimate. For the basic level, the averaging of the potential biases in the Angoff procedure does not reduce the potential problems because most of the panelists have large portions of their item ratings that are below their estimated standard.

The findings of the potential levels of biases are extremely interesting in light of the observed differences in the original cut scores set for these two methods in round 1 of standard setting. Schulz (2006) observed that the cut score estimate for Mapmark method was on average 6 points lower for the basic level, 20 points lower for the advanced level, and 6 points lower at the advanced level in the 2005 mathematics pilot study (Table 5.7). At the basic level, the overestimation of the Angoff cut scores and the underestimation of the Mapmark cut scores could potentially explain some of differences that are observed in applying these two standard setting methods since the sum of the potential biases are close to the difference in cut scores between the two methods.

Table 5.7: Difference Between Cut Score Estimates and Biases

Cut	Difference Between Cut Score of Bookmark and Angoff Methods	Potential Biases in Angoff	Potential Biases in Mapmark	Sum of Potential Biases
Basic	-6	-8.859	1.667	-7.192
Proficient	-20	-0.236	1.857	1.621
Advanced	-6	0.900	1.381	2.281

At the other two performance levels, the differences between the methods are harder to explain. It might be the case that panelists in the two standard settings differed significantly in interpreting the PLDs at the proficient and advanced levels and this translated into differences in the cut score estimates beyond the potential issues from not being able to accurately translate their intended cut score onto the score scale in their standard setting judgments. It might also be the case that additional issues arose in applying one of the standard setting procedures that could not be fully captured by the indices developed in this study. For example, panelists who applied the Mapmark method could have struggled with how to handle items being perceived to be out of order and how to make sense of this when providing standard setting judgments. Schulz (2006) noted that this was a potential problem when applying the Mapmark method in the pilot study. If panelists marked items too early in the Mapmark procedure for the proficient and advanced levels this would lower the cut score estimates for these two performance levels.

Unfortunately, it is impossible to detect problems of panelists marking items too early in their OIB or problems associated with items being perceived to be out of order with the indices developed in this dissertation. This is one potential limitation of the approach suggested in this dissertation. Additional discussion of this issue as well as

some of the limitations of the methods suggested in this dissertation is provided in the Chapter 6.

5.6 Practical Implications of Potential Biases

An important question to ask is what the impact of the potential biases in individual panelist and group cut score estimates could be on school accountability. Under NCLB, practical impact is often measured by the percentage of students that would change classification in different testing contexts. The changes in PAC between the estimated cut scores for the panelists or group of panelists and the estimated cut score plus the estimated biases are calculated based on the equations in Chapter 4 after rounding the panelist biases to the nearest whole number. The rounding of the biases to the nearest whole number is necessary since NAEP only reports student proficiency distributions at whole number scale score points ranging from 100 to 400 in the mathematics pilot study. The changes in the PAC for individual and group of panelists in the Angoff procedure are presented in Tables 5.8 and 5.9, respectively.

Table 5.8: Changes in PAC for Individual Panelists for Angoff Method

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
A1201	R1	Basic	235	-3	232	68.738	71.585	-2.847
A1202	R1	Basic	252	-11	241	50.379	62.603	-12.224
A1203	R1	Basic	255	-9	246	46.94	57.169	-10.229
A1204	R1	Basic	249	-13	236	53.771	67.798	-14.027
A1205	R1	Basic	267	3	270	33.106	29.679	3.427
A1206	R1	Basic	250	-5	245	52.695	58.306	-5.611
A1207	R1	Basic	256	-20	236	45.78	67.798	-22.018
A1208	R1	Basic	242	-18	224	61.538	78.313	-16.775
A1209	R1	Basic	245	-20	225	58.306	77.583	-19.277
A1210	R1	Basic	254	1	255	48.004	46.94	1.064
B1211	R1	Basic	242	-8	234	61.538	69.729	-8.191
B1212	R1	Basic	223	-15	208	79.086	88.829	-9.743
B1213	R1	Basic	245	-6	239	58.306	64.678	-6.372
B1214	R1	Basic	257	-5	252	44.533	50.379	-5.846
B1215	R1	Basic	239	-7	232	64.678	71.585	-6.907
B1216	R1	Basic	262	-9	253	38.701	49.209	-10.508
B1217	R1	Basic	250	-6	244	52.695	59.404	-6.709
B1218	R1	Basic	253	-4	249	49.209	53.771	-4.562
B1219	R1	Basic	233	-19	214	70.663	85.411	-14.748
B1220	R1	Basic	268	-3	265	31.887	35.353	-3.466
A1201	R2	Basic	245	-5	240	58.306	63.655	-5.349
A1202	R2	Basic	247	-5	242	56.006	61.538	-5.532
A1203	R2	Basic	256	-7	249	45.78	53.771	-7.991
A1204	R2	Basic	243	-8	235	60.472	68.738	-8.266
A1205	R2	Basic	241	2	243	62.603	60.472	2.131
A1206	R2	Basic	244	-1	243	59.404	60.472	-1.068
A1207	R2	Basic	252	-5	247	50.379	56.006	-5.627
A1208	R2	Basic	247	-2	245	56.006	58.306	-2.3
A1209	R2	Basic	246	-11	235	57.169	68.738	-11.569
A1210	R2	Basic	246	3	249	57.169	53.771	3.398
B1211	R2	Basic	245	-1	244	58.306	59.404	-1.098
B1212	R2	Basic	251	0	251	51.515	51.515	0
B1213	R2	Basic	244	-3	241	59.404	62.603	-3.199

Table 5.8 (cont'd)

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
B1214	R2	Basic	257	0	257	44.533	44.533	0
B1215	R2	Basic	226	-29	197	76.857	93.387	-16.53
B1216	R2	Basic	257	0	257	44.533	44.533	0
B1217	R2	Basic	243	-1	242	60.472	61.538	-1.066
B1218	R2	Basic	245	2	247	58.306	56.006	2.3
B1219	R2	Basic	226	-11	215	76.857	84.757	-7.9
B1220	R2	Basic	269	-2	267	30.778	33.106	-2.328
A1201	R3	Basic	245	-6	239	58.306	64.678	-6.372
A1202	R3	Basic	247	-6	241	56.006	62.603	-6.597
A1203	R3	Basic	249	-15	234	53.771	69.729	-15.958
A1204	R3	Basic	243	-7	236	60.472	67.798	-7.326
A1205	R3	Basic	247	2	249	56.006	53.771	2.235
A1206	R3	Basic	243	2	245	60.472	58.306	2.166
A1207	R3	Basic	252	-1	251	50.379	51.515	-1.136
A1208	R3	Basic	247	-4	243	56.006	60.472	-4.466
A1209	R3	Basic	246	-2	244	57.169	59.404	-2.235
A1210	R3	Basic	245	2	247	58.306	56.006	2.3
B1211	R3	Basic	243	0	243	60.472	60.472	0
B1212	R3	Basic	251	1	252	51.515	50.379	1.136
B1213	R3	Basic	241	-3	238	62.603	65.706	-3.103
B1214	R3	Basic	246	0	246	57.169	57.169	0
B1215	R3	Basic	230	-26	204	73.362	90.766	-17.404
B1216	R3	Basic	240	-7	233	63.655	70.663	-7.008
B1217	R3	Basic	239	-7	232	64.678	71.585	-6.907
B1219	R3	Basic	226	-7	219	76.857	82.068	-5.211
B1220	R3	Basic	270	-3	267	29.679	33.106	-3.427
A1201	R1	Proficient	271	-1	270	28.611	29.679	-1.068
A1202	R1	Proficient	294	7	301	9.412	6.116	3.296
A1203	R1	Proficient	284	0	284	16.131	16.131	0
A1204	R1	Proficient	295	-6	289	8.916	12.441	-3.525
A1205	R1	Proficient	287	10	297	13.801	7.913	5.888
A1206	R1	Proficient	289	2	291	12.441	11.148	1.293
A1207	R1	Proficient	293	2	295	9.976	8.916	1.06

Table 5.8 (cont'd)

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
A1208	R1	Proficient	280	-5	276	19.556	23.397	-3.841
A1209	R1	Proficient	273	-1	272	26.532	27.646	-1.114
A1210	R1	Proficient	285	10	295	15.338	8.916	6.422
B1211	R1	Proficient	275	2	277	24.479	22.376	2.103
B1212	R1	Proficient	273	-4	269	26.532	30.778	-4.246
B1213	R1	Proficient	289	-4	285	12.441	15.338	-2.897
B1214	R1	Proficient	285	-4	281	15.338	18.681	-3.343
B1215	R1	Proficient	282	-5	277	17.831	22.376	-4.545
B1216	R1	Proficient	302	3	305	5.713	4.608	1.105
B1217	R1	Proficient	287	-8	279	13.801	20.454	-6.653
B1218	R1	Proficient	286	5	291	14.62	11.148	3.472
B1219	R1	Proficient	291	-3	288	11.148	13.105	-1.957
B1220	R1	Proficient	292	-5	287	10.563	13.801	-3.238
A1201	R2	Proficient	281	-4	277	18.681	22.376	-3.695
A1202	R2	Proficient	290	5	295	11.721	8.916	2.805
A1203	R2	Proficient	286	0	286	14.62	14.62	0
A1204	R2	Proficient	282	-1	281	17.831	18.681	-0.85
A1205	R2	Proficient	268	7	275	31.887	24.479	7.408
A1206	R2	Proficient	270	3	273	29.679	26.532	3.147
A1207	R2	Proficient	286	2	288	14.62	13.105	1.515
A1208	R2	Proficient	281	0	281	18.681	18.681	0
A1209	R2	Proficient	274	0	274	25.433	25.433	0
A1210	R2	Proficient	278	8	286	21.385	14.62	6.765
B1211	R2	Proficient	279	1	280	20.454	19.556	0.898
B1212	R2	Proficient	284	-2	282	16.131	17.831	-1.7
B1213	R2	Proficient	289	-5	284	12.441	16.131	-3.69
B1214	R2	Proficient	287	-3	284	13.801	16.131	-2.33
B1215	R2	Proficient	268	-6	262	31.887	38.701	-6.814
B1216	R2	Proficient	297	-2	295	7.913	8.916	-1.003
B1217	R2	Proficient	280	-4	276	19.556	23.397	-3.841
B1218	R2	Proficient	280	-1	279	19.556	20.454	-0.898
B1219	R2	Proficient	288	-2	286	13.105	14.62	-1.515
B1220	R2	Proficient	294	-6	288	9.412	13.105	-3.693
A1201	R3	Proficient	281	-4	277	18.681	22.376	-3.695

Table 5.8 (cont'd)

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
A1202	R3	Proficient	289	4	293	12.441	9.976	2.465
A1203	R3	Proficient	283	-2	281	17.033	18.681	-1.648
A1204	R3	Proficient	282	-1	281	17.831	18.681	-0.85
A1205	R3	Proficient	272	6	278	27.646	21.385	6.261
A1206	R3	Proficient	267	2	269	33.106	30.778	2.328
A1207	R3	Proficient	287	1	288	13.801	13.105	0.696
A1208	R3	Proficient	281	0	281	18.681	18.681	0
A1209	R3	Proficient	280	1	281	19.556	18.681	0.875
A1210	R3	Proficient	278	7	285	21.385	15.338	6.047
B1211	R3	Proficient	276	2	278	23.397	21.385	2.012
B1212	R3	Proficient	285	-3	282	15.338	17.831	-2.493
B1213	R3	Proficient	283	-3	280	17.033	19.556	-2.523
B1214	R3	Proficient	276	-1	275	23.397	24.479	-1.082
B1215	R3	Proficient	271	-3	268	28.611	31.887	-3.276
B1216	R3	Proficient	278	0	278	21.385	21.385	0
B1217	R3	Proficient	278	-5	273	21.385	26.532	-5.147
B1218	R3	Proficient	263	1	264	37.57	36.442	1.128
B1219	R3	Proficient	287	-2	285	13.801	15.338	-1.537
B1220	R3	Proficient	295	-7	288	8.916	13.105	-4.189
A1201	R1	Advanced	301	-1	300	6.116	6.55	-0.434
A1202	R1	Advanced	330	6	336	0.458	0.212	0.246
A1203	R1	Advanced	321	-6	315	1.146	2.088	-0.942
A1204	R1	Advanced	334	-3	331	0.288	0.406	-0.118
A1205	R1	Advanced	308	18	326	3.719	0.692	3.027
A1206	R1	Advanced	310	9	319	3.136	1.407	1.729
A1207	R1	Advanced	332	3	335	0.358	0.251	0.107
A1208	R1	Advanced	311	-2	309	2.895	3.446	-0.551
A1209	R1	Advanced	297	10	307	7.913	3.998	3.915
A1210	R1	Advanced	317	10	327	1.717	0.632	1.085
B1211	R1	Advanced	297	3	300	7.913	6.55	1.363
B1212	R1	Advanced	318	2	320	1.544	1.273	0.271
B1213	R1	Advanced	328	-2	326	0.585	0.692	-0.107
B1214	R1	Advanced	309	-1	308	3.446	3.719	-0.273
B1215	R1	Advanced	317	-5	312	1.717	2.679	-0.962

Table 5.8 (cont'd)

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
B1216	R1	Advanced	337	-4	333	0.191	0.32	-0.129
B1217	R1	Advanced	321	-10	311	1.146	2.895	-1.749
B1218	R1	Advanced	330	0	330	0.458	0.458	0
B1219	R1	Advanced	311	-3	308	2.895	3.719	-0.824
B1220	R1	Advanced	323	-5	318	0.948	1.544	-0.596
A1201	R2	Advanced	312	-3	309	2.679	3.446	-0.767
A1202	R2	Advanced	323	11	334	0.948	0.288	0.66
A1203	R2	Advanced	328	0	328	0.585	0.585	0
A1204	R2	Advanced	306	10	316	4.337	1.885	2.452
A1205	R2	Advanced	282	18	300	17.831	6.55	11.281
A1206	R2	Advanced	294	6	300	9.412	6.55	2.862
A1207	R2	Advanced	323	5	328	0.948	0.585	0.363
A1208	R2	Advanced	309	1	310	3.446	3.136	0.31
A1209	R2	Advanced	299	3	302	6.965	5.713	1.252
A1210	R2	Advanced	300	8	308	6.55	3.719	2.831
B1211	R2	Advanced	300	2	302	6.55	5.713	0.837
B1212	R2	Advanced	317	1	318	1.717	1.544	0.173
B1213	R2	Advanced	326	0	326	0.692	0.692	0
B1214	R2	Advanced	314	-2	312	2.285	2.679	-0.394
B1215	R2	Advanced	299	2	301	6.965	6.116	0.849
B1216	R2	Advanced	329	-5	324	0.53	0.846	-0.316
B1217	R2	Advanced	313	-4	309	2.472	3.446	-0.974
B1218	R2	Advanced	312	-1	311	2.679	2.895	-0.216
B1219	R2	Advanced	309	-1	308	3.446	3.719	-0.273
B1220	R2	Advanced	324	-5	319	0.846	1.407	-0.561
A1201	R3	Advanced	312	-3	309	2.679	3.446	-0.767
A1202	R3	Advanced	318	13	331	1.544	0.406	1.138
A1203	R3	Advanced	319	0	319	1.407	1.407	0
A1204	R3	Advanced	306	11	317	4.337	1.717	2.62
A1205	R3	Advanced	288	14	302	13.105	5.713	7.392
A1206	R3	Advanced	288	3	291	13.105	11.148	1.957
A1207	R3	Advanced	321	6	327	1.146	0.632	0.514
A1208	R3	Advanced	309	2	311	3.446	2.895	0.551
A1209	R3	Advanced	306	4	310	4.337	3.136	1.201

Table 5.8 (cont'd)

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
A1210	R3	Advanced	300	5	305	6.55	4.608	1.942
B1211	R3	Advanced	298	3	301	7.383	6.116	1.267
B1212	R3	Advanced	316	0	316	1.885	1.885	0
B1213	R3	Advanced	308	9	317	3.719	1.717	2.002
B1214	R3	Advanced	308	-2	306	3.719	4.337	-0.618
B1215	R3	Advanced	303	3	306	5.334	4.337	0.997
B1216	R3	Advanced	308	-3	305	3.719	4.608	-0.889
B1217	R3	Advanced	311	-6	305	2.895	4.608	-1.713
B1218	R3	Advanced	299	-2	297	6.965	7.913	-0.948
B1219	R3	Advanced	309	-1	308	3.446	3.719	-0.273
B1220	R3	Advanced	325	-6	319	0.776	1.407	-0.631

These results suggest that there is the potential for the biases in the individual panelist ratings to translate into large changes in the PAC for individual panelists (Table 5.8). At the basic level, the changes in PAC can be a decrease of as much as 22 percent of students being classified as being above the basic cut score for an individual panelist in the first round for the estimated cut score compared to the cut score based on the individual item ratings. At the other cut score placements, the magnitude of the changes is often not as severe, but can still be somewhat extreme in the current context of NCLB. For example, many panelists would see changes in student classification rates ranging from 3 to 10 percent. These changes in classification rates at the different performance levels are very high compared to the changes that are commonly observed in many state and national accountability programs.

Table 5.9: Changes in PAC for Group of Panelists for Angoff Method

Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
R1	Basic	249	-9	240	53.771	63.655	-9.884
R2	Basic	247	-4	243	56.006	60.472	-4.466
R3	Basic	244	-5	239	59.404	64.678	-5.274
R1	Proficient	286	0	286	14.62	14.62	0
R2	Proficient	282	0	282	17.831	17.831	0
R3	Proficient	280	0	280	19.556	19.556	0
R1	Advanced	318	1	319	1.544	1.407	0.137
R2	Advanced	311	2	313	2.895	2.472	0.423
R3	Advanced	308	2	310	3.719	3.136	0.583

For the group of panelists, the potential changes in the PAC are large for the basic level across the three rounds of standard setting judgments and minimal for the other two performance levels. For the basic level, the changes in PAC range from -4.426 to -9.884. This means that the number of students classified as being above the basic level is lower than expected if the level of the aggregate potential bias of the panelists is considered. These levels of students changing classification at the basic level are again extremely large in the context of NCLB. For the other two performance levels, the degree of potential changes in the PAC might be viewed as somewhat encouraging, especially at the proficient level where there is no change in the PAC given that the proficient level is often the important cut score used for measuring AYP in most state testing programs. However, the essential realization from comparing Tables 5.8 and 5.9 is that panelist bias does have the potential to translate in changes in the PAC that are practically significant at the group level.

Similar to the findings for the Angoff method presented in Tables 5.8 and 5.9, the biases in the Mapmark procedure can also be converted in changes in the PAC at the

individual and group levels. The Mapmark changes in PAC are displayed in Tables 5.10 and 5.11.

Table 5.10: Changes in PAC for Individual Panelists for Mapmark Method

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
MA1201	R1	Basic	214	1	215	85.411	84.757	0.654
MA1202	R1	Basic	226	1	227	76.857	76.093	0.764
MA1203	R1	Basic	237	1	238	66.778	65.706	1.072
MA1204	R1	Basic	212	2	214	86.672	85.411	1.261
MA1205	R1	Basic	227	3	230	76.093	73.362	2.731
MA1206	R1	Basic	246	1	247	57.169	56.006	1.163
MA1207	R1	Basic	246	1	247	57.169	56.006	1.163
MA1208	R1	Basic	257	2	259	44.533	42.21	2.323
MA1209	R1	Basic	246	1	247	57.169	56.006	1.163
MA1210	R1	Basic	245	1	246	58.306	57.169	1.137
MB1211	R1	Basic	254	1	255	48.004	46.94	1.064
MB1212	R1	Basic	230	3	233	73.362	70.663	2.699
MB1213	R1	Basic	243	2	245	60.472	58.306	2.166
MB1214	R1	Basic	255	2	257	46.94	44.533	2.407
MB1215	R1	Basic	252	1	253	50.379	49.209	1.17
MB1216	R1	Basic	246	2	248	57.169	54.845	2.324
MB1217	R1	Basic	213	2	215	86.048	84.757	1.291
MB1218	R1	Basic	255	2	257	46.94	44.533	2.407
MB1219	R1	Basic	225	1	226	77.583	76.857	0.726
MB1220	R1	Basic	213	2	215	86.048	84.757	1.291
MB1221	R1	Basic	239	3	242	64.678	61.538	3.14
MA1201	R1	Proficient	251	1	252	51.515	50.379	1.136
MA1202	R1	Proficient	270	3	273	29.679	26.532	3.147
MA1203	R1	Proficient	273	1	274	26.532	25.433	1.099
MA1204	R1	Proficient	247	4	251	56.006	51.515	4.491
MA1205	R1	Proficient	247	4	251	56.006	51.515	4.491
MA1206	R1	Proficient	266	1	267	34.215	33.106	1.109

Table 5.10 (cont'd)

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
MA1207	R1	Proficient	266	1	267	34.215	33.106	1.109
MA1208	R1	Proficient	276	1	277	23.397	22.376	1.021
MA1209	R1	Proficient	270	3	273	29.679	26.532	3.147
MA1210	R1	Proficient	270	3	273	29.679	26.532	3.147
MB1211	R1	Proficient	285	1	286	15.338	14.62	0.718
MB1212	R1	Proficient	257	2	259	44.533	42.21	2.323
MB1213	R1	Proficient	265	1	266	35.353	34.215	1.138
MB1214	R1	Proficient	291	1	292	11.148	10.563	0.585
MB1215	R1	Proficient	275	2	277	24.479	22.376	2.103
MB1216	R1	Proficient	255	2	257	46.94	44.533	2.407
MB1217	R1	Proficient	257	2	259	44.533	42.21	2.323
MB1218	R1	Proficient	287	1	288	13.801	13.105	0.696
MB1219	R1	Proficient	265	1	266	35.353	34.215	1.138
MB1220	R1	Proficient	257	2	259	44.533	42.21	2.323
MB1221	R1	Proficient	257	2	259	44.533	42.21	2.323
MA1201	R1	Advanced	294	1	295	9.412	8.916	0.496
MA1202	R1	Advanced	307	1	308	3.998	3.719	0.279
MA1203	R1	Advanced	312	2	314	2.679	2.285	0.394
MA1204	R1	Advanced	295	2	297	8.916	7.913	1.003
MA1205	R1	Advanced	307	1	308	3.998	3.719	0.279
MA1206	R1	Advanced	326	2	328	0.692	0.585	0.107
MA1207	R1	Advanced	316	1	317	1.885	1.717	0.168
MA1208	R1	Advanced	317	3	320	1.717	1.273	0.444
MA1209	R1	Advanced	316	1	317	1.885	1.717	0.168
MA1210	R1	Advanced	316	1	317	1.885	1.717	0.168
MB1211	R1	Advanced	322	2	324	1.032	0.846	0.186
MB1212	R1	Advanced	299	1	300	6.965	6.55	0.415
MB1213	R1	Advanced	321	1	322	1.146	1.032	0.114
MB1214	R1	Advanced	321	1	322	1.146	1.032	0.114
MB1215	R1	Advanced	321	1	322	1.146	1.032	0.114
MB1216	R1	Advanced	301	1	302	6.116	5.713	0.403
MB1217	R1	Advanced	312	1	313	2.679	2.472	0.207

Table 5.10 (cont'd)

Rater	Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
MB1218	R1	Advanced	316	1	317	1.885	1.717	0.168
MB1219	R1	Advanced	305	3	308	4.608	3.719	0.889
MB1220	R1	Advanced	312	1	313	2.679	2.472	0.207
MB1221	R1	Advanced	302	1	303	5.713	5.334	0.379

Compared to the results for the Angoff method, the changes in the PAC for individual panelists tend to be somewhat smaller in magnitude (Table 5.10). For example, there is not a single individual panelist where the PAC would change by more than five percent. At the advanced level, the changes from considering the maximum potential bias in cut scores are always less than one percent for all panelists. However, there are several panelists where the change in PAC could be as large as three to four percent and this level of potential changes in student classification might be viewed as being a concern by some policymakers. These greater potential changes in student classification tend to occur at the basic level since the density of students is greater in the regions where the basic level cut score is being placed compared to the other proficiency levels. This means that the same change in the cut score estimate at this performance level could turn into a greater change in the PAC. This also implies that the same level of absolute bias can be more or less of a concern in terms of the PAC depending on how many students are in the region where the cut score is located

Table 5.11: Changes in PAC for Group of Panelists for Mapmark Method

Round	Cut	Estimate	Bias	Estimate + Bias	PAC Estimate	PAC Estimate + Bias	Change in PAC
R1	Basic	243	2	245	60.472	58.306	2.166
R1	Proficient	266	2	268	34.215	31.887	2.328
R1	Advanced	312	1	313	2.679	2.472	0.207

At the group level, the changes in PAC can range from 0.207 percent at the advanced level to 2.166 percent at the basic level to 2.328 percent at the proficient level for the Bookmark procedure (Table 5.11). These changes in PAC are always positive indicating that there is the potential for percentage of students to be overestimated when applying the Mapmark method. In comparison to the Angoff procedure, the absolute magnitude of the potential changes in cut scores could be larger for the Mapmark procedure at the proficient and advanced levels and less at the basic level. Again, even these small changes in PAC might be considered a concern for policymakers, especially if the true changes in the PAC are masked by potential biases in the intended cut scores of the standard setting panelists.

5.7 Discussion of Empirical Comparisons

From a theoretical perspective the results presented in Sections 5.2 through 5.6 are compelling, but not completely unexpected. Shephard et al. (1993) have pointed out that panelists struggle to rate consistently with a specific level of test performance when applying the Angoff method and Schulz (2006) suggested that this might be an issue in the 2005 mathematics pilot study. What is interesting is the fact that rater inconsistency in the Angoff procedure has the potential to produce high levels of bias for both an

individual and groups of panelists. It is important to point out, however, that not every panelist or groups of panelists will have high levels of bias. This is evidenced by the low levels of potential bias for the group of panelists at the proficient level in the NAEP data.

These levels of potential biases have been previously undiscovered and have not been reported in any operational study involving the Angoff procedure to date. These findings of potentially high biases seem to substantiate the common observations made in the literature (Shepard, et al., 1993; Impara & Plake, 1997; Linn & Shepard, 1997; Schulz, 2005) that rater inconsistency in the Angoff procedure can lead to problems when applying the Angoff method. The high levels of potential bias for an individual panelist when applying the Angoff procedure support the suggestion to not rely on a single or small number of panelists' judgments in standard setting since there is the possibility for the individual panelists' judgments to exhibit high levels of bias.

Another essential observation in the context of the results presented for the Angoff procedure are the levels of inconsistency reported in various performance categories across rounds of the standard setting process. For example, a general pattern observed in fixed effects models with the absolute residuals was that the estimated coefficients tended to decrease across rounds. These general patterns of increased consistency are not entirely surprising when applying the Angoff procedure used in the NAEP pilot study since panelists receive Reckase charts, which in theory show panelists how to provide judgments that are in line with any specific level of test performance. The fact that the coefficients for the rater inconsistency are not equal to zero after receiving the Reckase charts as feedback means that panelists either do not fully understand how to

use the Reckase charts or that panelists still choose to give inconsistent ratings when presented with this type of feedback.

The finding of decreased inconsistency over rounds of the standard setting process would appear to support the suggestion that feedback between rounds of the standard setting process can help panelists to refine their standard setting ratings and be more accurate in later rounds of standard setting. It might also suggest that a single round of Angoff standard setting is not desirable since panelists have the potential to be very inconsistent when only a single round is used. It is important to point out, however, that more consistent ratings may not necessarily lead to less bias in standard setting judgments.

The finding that the basic level had the most inconsistent judgments in the Angoff procedure can be explained by examining the item rating data more closely. As others have mentioned previously in the research literature (Reckase, 2001), panelists often choose to give probability ratings of items at the lowest performance level that are below the guessing parameter of the items. For example, several of the 3PL items used in the Angoff rating activity had a *c*-parameter of around 0.3. Several of the panelists provided ratings that were lower than this *c*-parameter in their initial standard setting judgments. These low item ratings translate into larger amounts of bias and greater inconsistency at this performance level.

The findings of potential negative biases and inflated cut scores when applying the Mapmark method supports the idea postulated by Reckase (2006a) in his simulation study. Specifically, Reckase (2006a) suggested that the item difficulty gaps in the Mapmark method have the potential to result in cut scores that are underestimated for

individual panelists. The empirical investigations of the Mapmark procedure in the mathematics pilot study suggest that these item difficulty gaps have the potential to generate biases in cut score estimates that might be practically significant. The extension of investigating the potential bias at the group level shows that not only is there the potential for these item difficulty gaps to cause problems for individual panelists, they can also present issues for a group of panelists. Further, since these item difficulty gaps always have the potential to result in cut scores that are underestimated for each panelist it is not possible for the biases to cancel out for a group of panelists. This stands in direct contrast to the Angoff procedure where it is possible for the panelists' biases to cancel out. However, this does not suggest that the Angoff method is universally preferred over the Bookmark procedure. The potential biases observed when applying these methods in practice are a function of the ratings provided by the panelist and the cut score that they want to specify. This suggests that either procedure can have higher or lower levels of bias in practice depending on the ratings, the item used in standard setting, and the intended cut scores.

Finally, the investigations of how the potential biases in the Angoff and Mapmark procedures would impact the PAC provide key insights into what the potential biases could mean in terms of classifying students in a large scale assessment program. In this context, the findings presented in Section 5.7 suggest that there is the potential for practically significant changes in the PAC in both the Angoff and Bookmark standard setting procedures for panelist's inability to accurately indicate intended cut scores. Given the high stakes associated with these statistics and the fact that cut score estimates might be set higher or lower than intended this is an area in need of more research.

CHAPTER 6

CONCLUSION

The purpose of this dissertation is to introduce a new comprehensive framework that could be used to evaluate standard setting methods. This framework was built on previous suggestions of Reckase (2006a; 2006b) and construct maps (Wilson, 2005). The framework emphasizes the desire for panelists to arrive at a hypothetical intended cut score in standard setting. The new framework extends previous research by not only formulating how evaluations can be performed for an individual panelist in simulated situations, but also showing how the framework can be applied to evaluate standard setting judgments in both operational and simulated settings for individual panelists or groups of panelists. Examples of how to apply the new framework to develop indices to evaluate and compare the two most commonly applied standard setting methods, the Angoff and Bookmark methods, were provided. These examples included the evaluation of both the Angoff and Mapmark (a variation of Bookmark) methods using newly developed indices for the 2005 mathematics pilot study of NAEP.

Results from the investigations of the 2005 mathematics pilot study data suggested that there is the potential for the cut scores at both the individual and group level to be impacted by potential biases and inconsistencies. Specifically, the indices developed based on the new integrated framework showed that there is the potential for the Angoff procedure to be adversely impacted by rater inconsistency, which could bias cut score estimates and change the percentage of students classified into the different performance categories. Potential issues with the Mapmark procedure from item

difficulty gaps between items were also identified in the 2005 mathematics pilot study of NAEP. In particular, the presence of item difficulty gaps in the OIB used with the Mapmark procedure were shown to have the potential to result in the estimated cut scores that were lower than intended. These lower than intended cut scores estimates in turn could result in the percentage of students classified at the different performance levels be higher than they should.

6.1 Unique Contributions

These analyses and the newly developed indices highlight the capacity of the new framework to inform and improve standard setting practices. Specifically, it helps researchers and practitioners with conceptualizing and evaluating standard setting judgments from a new perspective that focuses on the link between the cut score that a panelist intends to set (a hypothetical cut score that a panelist had in mind when they provided their standard setting judgments) and how accurate they are at setting this cut score. This important question often drives most evaluations of standard setting methods – “How well do the standard setting judgments represent the cut scores that the panelists’ intended?”

By conceptualizing standard setting through the lens of this framework the aforementioned question can be directly investigated and answered in any situation in which standard setting might be performed. For example, if one were interested in knowing how a range of standard setting judgments might impact the cut score estimates for the Bookmark procedure for a particular set of test items, it is possible to use the new framework to investigate what levels of potential bias might be present for these range of

judgments for this set of test items. These types of investigations using the framework prior to operational standard setting can help to identify potential concerns that might arise in standard setting and can lead to potential remedies that could reduce the levels of bias observed in operational situations. For example, if it is found that the item difficulty gaps in the regions where the cut score is to be estimated in Bookmark method could lead to undesirable levels of bias, more test items can be added in these regions to reduce the item difficulty gaps between the items. In conjunction with the ability to use the framework to develop evaluation indices that can be applied in operational standard setting situations, this framework has the potential to help reduce the impact of biases and inconsistencies on cut score estimates.

Another potential application of the new framework to improve standard setting would be to use the evaluation indices after each round of standard setting judgments as feedback mechanisms to decrease panelist bias and inconsistency. For example, in NAEP standard setting that uses the Angoff procedure with Reckase charts, it is possible to calculate the coefficients of the fixed effects models applied in this dissertation and to present this information along side the Reckase charts as feedback. This information could help cement in panelists' minds the potential impact that not providing standard setting judgments in line with a specific level of test performance could have on estimated cut scores. One issue that has continued to plague standard setting has been the difficulty of panelists to see the relationship between the judgments that they provide in standard setting and how these judgments relate to the estimated cut score. By using the indices and construct maps as part of standard setting this makes it clearer to panelists how all of these quantities are related. Of course, this might change the way that panelists

provide their standard setting judgments as well as panelists understanding of what it means to set cut scores on assessments.

This brings forth yet another unique contribution presented in this dissertation; the idea of expanding the concept of a construct map to include all of the information that exists within an IRT framework as shown in Table 3.1 and illustrating how it might be used in the context of standard setting. This expanded notion of construct maps is essentially an extension of the work of Schulz and Mitzel (2005; in press) who have begun to use construct maps that contain domain score feedback in an attempt to improve standard setting. The essential realization is that it is possible to relate not only item locations and domain score feedback together in a construct map, but it is also possible to include information on work samples, the PAC, the performance of students in teacher's classrooms, and a host of other quantities. The linking of all of this information together in a tabular format is designed to help both standard setters and panelists to have a better understand of the many relationships that exist between quantities which are often hard to relate to each other in a simple fashion in one's mind.

There are many ways in which the construct map could be used in conjunction with the evaluation indices not only as feedback mechanisms to improve standard setting, but also to develop new standard setting procedures that might lead to better standard setting judgments. For example, it is possible to have panelists select a profile of expected performance in the content domains in a construct map as their cut score estimate as a round of standard setting. Then, in subsequent rounds additional information could be added to the construct map (i.e., whole booklets, PAC, etc.,) and this information in conjunction with the domain score information could be used to guide

estimation of cut scores. It is also possible to combine the Angoff and Bookmark standard setting methods together to get a hybrid standard setting method using construct maps. It is expected that future research could produce several new improved standard setting methods based on construct maps.

Another unique outgrowth of the framework proposed in this dissertation is the ability to identify concerns in common standard setting methods that present threats to producing unbiased cut score estimates. In particular, the framework highlights that the reasons that cut score estimates might be biased is because of problems of rater inconsistency, gaps in score locations from the lack of stimulus at specific score scale locations, or from a combination of these two problems. Although some researchers have suggested that these issues might presents threats to the cut scores that are estimated, a direct relationship between these threats and the cut scores that are estimated is lacking in the literature. The limited understanding of these threats and how they relate to cut score estimates can probably best be explained by the fact that until this dissertation the notion of construct maps and the relationship between standard setting judgments and cut score estimates was often some what of a black box (McGinty, 2005).

6.2 Limitations and Concerns

One of the biggest concerns in the standard setting literature has revolved around the apparent arbitrary nature of cut scores that are set (Glass, 1978; Hambleton, 1978; Popham, 1978). In a classic critique of standard setting, Glass (1978) suggested that there is no true cut score estimate and that all cut scores are in some sense arbitrary. The framework presented in this dissertation presents a different view on this classic debate.

The view presented in this dissertation is that it is not the cut score and the ratings that are arbitrary, but rather it is the definition of the standard as outlined in the written description that is often arbitrary in the sense that it is impossible to define a true representation of the standard as expressed in the PLD. To the person that is not familiar with standard setting, the distinction between a true representation of a cut score and a true representation of standard may not seem that important. But in reality, this distinction is extremely important because it cuts to the heart of the real issue in standard setting: is it possible set a cut score that represents the PLD?

The important realization is that in many practical situations the standards and the written descriptions of these standards are often designed without explicitly considering the difficulty of the test items or the progression of learning that examinees gain in different content areas. Instead, the PLDs are often general statements of what content experts and policy makers think students should know at different levels, which often can be somewhat disconnected from test development and may not be systematically related to the knowledge, skills, and abilities that students acquire at different levels of the score scale that underlies the assessment. This lack of clarity in the definition of the PLDs as well as the limited relationship to actual test performance leads to ambiguity in the PLDs and it means that it is usually not possible to define a true representation of the PLD.

True representations of a cut score can be defined, however, since a cut score is a number on a score scale. For example, one can define a cut score to be 180 on a score scale that ranges from 100 to 400 and one can determine if a panelist who intended to set their cut score at 180 was able to set their cut score at this level by examining the ratings provided by the panelists.

The framework developed in this dissertation does not address the question of whether the intended cut score of a panelist or group of panelists is aligned with what the policy board intended when it defined the PLD. It also does not address the question of whether it is even possible to provide a cut score estimate that is in line with this PLD. It is also important to observe that neither Kane's (Kane, 2001) validity framework nor the Engelhard's (in press) MRM evaluation approach provide a direct answer to this question either because the relationship between the PLDs and cut scores is not explicitly considered in these evaluation approaches.

The reason this might be viewed as a limitation of the framework developed in this dissertation is because the framework cannot address problems of whether the cut score estimates provided by the panelists are meaningful in relationship to the PLD. To date, a limited amount of research has been performed examining the relationship between PLDs, test items, and cut scores with the notable exceptions of recent papers by Schneider et al. (2009), Ferrera et al. (2009), and Plake et al. (2009). Additional research that looks more clearly at these relationships would be very valuable to improving standard setting as well as increasing the meaning of performance levels in score reporting.

This also means that certain issues that can arise in standard setting from other types of biases that panelists might bring to bear in their standard setting judgments are very difficult to identify using the framework developed in this dissertation. For example, in the current high stakes testing situation it might be the case that panelists lower their cut score estimates across rounds of standard setting in response to discussion and feedback in the standard setting process so that the cut scores that are estimated are not

representative of the PLDs. That is, panelists may intend to set their cut score at a lower level than was intended in the PLD because it is advantageous to do so. For example, teacher's may lower cut scores because they may realize that if the cut score is set at a higher level it might mean increased pressure or other possible ramifications for them or the school that they work at in the future. This increased pressure or other possible ramifications may be undesirable meaning that it is to their advantage to lower the cut scores. The author is aware of at least one situation where this very issue arose in a standard setting after panelists were made aware of the percentage of students that would be classified at various performance levels. In this case, panelists chose to lower their cut score estimates not because their cut score estimates were necessarily inaccurate, but instead because the panelists were concerned about the potential ramifications of the cut scores for school accountability and teacher evaluation. Issues of panelists bookmarking items too early in the Bookmark procedure as well as having difficulty determining a cut score from items being perceived to be out of order in an OIB fall into the same class of issues as lowering the standard in response to student performance data. These concerns are also difficult to tease out from a single cut score estimate.

Unfortunately, these types of concerns are often more nuanced and require different analytical approaches beyond those presented in this dissertation or in the research literature. For example, observation of the standard setting process and listening to comments and conversations that take place during the standard setting meeting can provide clues to potential problems that are not fully captured by the statistical indices proposed in this dissertation or the responses provided on evaluation forms. Oftentimes, this type of information can go undocumented and in some circumstances may not be

presented to the policy board as they deliberate in considering what cut scores to adopt. More careful attention to these issues in standard setting is extremely important in understanding aspects of standard setting that are often difficult to statistically measure.

An additional limitation of the framework developed in this dissertation and the collection of standard setting ratings in general might be the apparent restrictive assumption of assuming that panelists provide independent ratings. In many circumstances, the tenability of this assumption is questionable given that panelists are allowed to interact with each other and discuss their standard setting judgments. However, this assumption is almost always made in practice and is usually employed in the procedures that are used to aggregate panelists' individual cut score estimates and for calculating standard errors. The impact of the possible interdependencies among raters and rounds of standard setting are unknown. Similarly, how these interdependencies impact the indices developed in this dissertation is also unknown. Additional research exploring these issues would also be quite valuable in the future.

6.3 Future Research

There are many areas for future research that are related to the work presented in this dissertation. Specifically, the comparisons and analyses performed in this dissertation were for two particular standard setting approaches on one set of standard setting data. Additional research that looks at how the new framework could be applied to other standard setting methods and in other situations would be quite valuable. For example, it would be useful to apply the Angoff and Bookmark indices to other tests and standard setting situations since many other testing programs use smaller samples of items and less

extensive training processes than are used in NAEP. The issue of the length of the tests presented to panelists to provide their standard setting judgments is an especially important area for future research. Currently, very little research has been performed that looks at how test length and the characteristics of items presented to standard setters impacts the cut scores that are set on large scale assessments. It might be the case that rater inconsistency in the Angoff procedure and the item difficulty gaps in the Bookmark method present bigger concerns in these other testing situations. For example, if a smaller sample of items is used in the Bookmark procedure the item difficulty gaps in the regions where panelists intend to set their cut score might become exaggerated, which could result in larger potential biases in cut score estimates. Obviously, more research is needed into how the characteristics of the test items and the sample of items used in standard setting impacts cut scores.

Another area for future research would be to apply the new standard setting framework to create new standard setting methods based on the construct maps. For example, new standard setting methods could be developed that use the construct maps and emphasize different components in the construct mapping framework in different rounds of the standard setting process. The development of these new standard setting methods could help to improve standard setting practices in circumstances in which setting standards have been challenging. In particular, it would be possible to use the construct mapping framework to develop new standard setting methods to determine cut scores across various grades levels if tests could be placed onto a vertical scale using IRT methods since cross-grade performance could be related together in a construct map. It is also would be possible to combine features of different standard setting methods together

using construct maps, so that the strengths of different methods can be realized when determining cut scores. For example, it would be possible to create hybrid standard setting methods in which panelists performed Bookmark and Angoff type standard setting judgments in different rounds of standard setting.

Additional research is also needed on how the novel indices and framework could be used as feedback between rounds of standard setting. How would panelists respond to the construct maps and the indices that are presented to them? Which types of feedback do panelists respond most positively to? In what circumstances would panelists provide ratings that are completely consistent with their intended cut scores? Relatively little is known about how panelists respond to feedback in general and even less is known about how panelists would respond to the construct maps suggested in this dissertation. Little is also known about the complex dependencies that can exist between raters after receiving feedback and how this impacts standard setting judgments. Coming up with methods for controlling and detecting these dependencies in standard setting would also be a useful direction for additional research.

Finally, more research is needed to resolve the apparent disconnect between PLDs and cut scores. The framework proposed in this dissertation could provide one avenue with which to begin to investigate these issues since the construct maps (see Table 3.1) could be modified and used to help content experts and policy makers when they construct and write the PLDs. For example, experts who are asked to write PLDs could consider the information in the construct map as they think about what students at different performance levels should know and be able to do. These construct maps could help experts to better see the relationship between test content and the construct that is

measured by the test items, which is often hard to conceptualize in practice. If these relationships were considered and accounted for when the PLDs were constructed this would make the PLDs more meaningful and provide a link between the written description in the PLDs and the skills, knowledge, and abilities required on the assessment. This relationship between the PLDs used for score reporting and cut scores set on the assessment is essential to ensuring the validity of test score interpretations.

REFERENCES

- ACT, Inc. (2005, April). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Special Studies report*. Iowa City, IA: Author.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Anderson, E. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69-81.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-597). Washington, DC: American Council on Education.
- Bay, L. (1998, March). *1998 NAEP achievement levels-setting process field trial 1 for civics*. Paper presented at the meeting of the Technical Advisory Committee for Standard Setting, Chicago.
- Beretvas, N. S. (2004). Comparison of Bookmark difficulty locations under different item response models. *Applied Psychological Measurement*, 28, 25-47.
- Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before). *Applied Measurement in Education*, 9, 215-235.
- Berk, R. A. (1986). A consumer's guide to setting performance on criterion-referenced tests. *Review of Educational Research*, 56, 137-172.
- Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45, 4-9.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wiley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in tow or more nominal categories. *Psychometrika*, 37, 29-51.
- Bourque, M.L. and Byrd, S. (Eds.) (2000). *Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements*. Washington, DC: National Assessment Governing Board.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard setting topics. *Applied Measurement in Education*, 17, 59-88.

- Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*, 253-263.
- Caines, J., & Engelhard, G. (2009, April). *Evaluating body of work judgments of standard-setting panelists*. Paper presented at the annual meeting of the American Educational Research Association. San Diego, CA.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications, Thousand Oaks: CA.
- Cizek, G. J. (2001). *Setting performance standards*. Mahwah, NJ: Lawrence Erlbaum.
- Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice, 15*, 12-21.
- Council of Chief State School Officers (2001). *State student assessment program annual survey. Data volume II*. Washington, DC: Author.
- Davis, S. L., Buckendahl, C. W., Chin, T., & Gerrow, J. (2008, March). *Comparing the Angoff and Bookmark methods for an international licensure exam*. Paper presented at the annual meeting of the National Council of Measurement in Education. New York, NY.
- Engelhard, G. (in press). Evaluating the judgments of standard-setting panelists using Rasch Measurement Theory. In E. V. Smith, Jr., and G. E. Stone (Eds.), *Applications of Rasch measurement in criterion-referenced testing*, JAM Press.
- Engelhard, G. (2007). Evaluating Bookmark judgments. *Rasch measurement transactions, 21*, 1097-1098.
- Engelhard, G., & Anderson, D. W. (1998). A binomial trials model for examining the ratings of standard-setting judges. *Applied Measurement in Education, 11*, 209-230.
- Engelhard, G., and Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement, 58*, 179-196.
- Ferrera, S. F., Svetina, D., Skucha, S. M., & Murphy, A. (2009, April). *Test development with standard setting and growth in mind*. Paper presented at the annual meeting of the National Council for Measurement in Education. San Diego, CA.
- Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research, 59*, 315-328.

- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.
- Goertz, M.E. (2001, September). *The federal role in defining "adequate yearly progress:" The flexibility/accountability trade-off*. Consortium for Policy Research in Education. Retrieved July 10, 2009, from http://www.cpre.org/images/stories/cpre_pdfs/cep01.pdf
- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Hambleton, R. K. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In Hansche, L.N. (Ed.). *Handbook for the development of performance standards: Meeting the requirements of Title I*. Bethesda, MD: Frost Associates.
- Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 15, 277-290.
- Hambleton, R. K. & Pitoniak (2006). Setting performance standards. In R. L. Brennan (Ed.) *Educational measurement* (4th edition), pp. 433-470. Washington, DC: American Council on Education.
- 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., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research & Perspective*, 2, 61-103.
- Hurtz, G. M., & Hertz, N. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement*, 59, 885-897.
- Hurtz, G. M., & Jones, J. P. (2009). Innovations in measuring rater accuracy in standard setting: Assessing "fit" to item characteristic curves. *Applied Measurement in Education*, 22, 120-143.

- Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice*, 25, 19-20.
- Improving America's Schools Act of 1994, Pub. L. No. 103-382. (1994). Retrieved September 9, 2009, from <http://www.ed.gov/legislation/ESEA/toc.html>
- Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 23, 35-56.
- Impara, J. C. & Plake, B. S. (1998). Teacher's ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 34, 353-366.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.) *Education measurement* (3rd edition), pp. 485-514. New York: American Council of Education and Macmillan.
- Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis*, 4, 461-475.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational measurement* (4th edition), pp. 17-64. Washington, DC: American Council on Education.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Kane, M. (1998). Criterion bias in examinee-centered standard setting: Some thought experiments. *Educational Measurement: Issues and Practice*, 17, 23-30.
- Kane, M. (1995). Examinee-centered vs. test-centered standard setting. In *Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II* (pp. 119-141). Washington, DC: U. S. Government Printing Office.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.
- Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527-535.
- Kane, M. T. (1987). On the use of IRT models with judgmental standard-setting procedures. *Journal of Educational Measurement*, 24, 333-345.

- Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 26, 4-12.
- Kolstad, A. (2002, June). *Various approaches to providing content-referenced interpretations for IRT scale reporting: NAEP's anchor levels, adult literacy levels, and PISA levels*. Paper prepared for presented to the National Conference on Large-Scale Assessment. Palm Desert, CA.
- Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (2001) *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Working Paper No. 2001-20. Washington, DC: National Center for Education Statistics.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.), pp.531-578. Westport, CT: American Council on Education/ Praeger.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium presentation at the Council of Chief State School Officers National Conference of Large-Scale Assessment, Phoenix, AZ.
- Linn, R. L. (2003a). Accountability: Responsibility and reasonable expectations. 2003 presidential address. *Educational Researcher*, 32, 3-17.
- Linn, R. L. (2003b). Performance standards: Utility for different uses of assessments. *Educational Policy Analysis Archives*, 11(31), Retrived January, 26, 2008, from <http://epaa.asu.edu/epaa.v11n31>.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31, 3-16.
- Linn, R. L., & Shepard, L. A. (1997, July). *Item-by-item standard setting: Misinterpretations of judge's intentions due to less than perfect item inter-correlation*. Paper presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Colorado Springs, CO.
- 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.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Masters, G. N., Adams, R., & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research*, 21, 595-609.
- McGinty, D. (2005). Illuminating the “black box” of standard setting: An exploratory qualitative study. *Applied Measurement in Education*, 18, 269-287.
- Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II* (pp. 221-263). Washington, DC: U. S. Government Printing Office.
- Messick, S. (1989). Validity. In R. L. Linn (Ed), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- Messick, S. (1988). The once and future issues of validity. Assessing and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E. & Bock, R. D. (1991). *PARSCALE: Parameter scaling of rating data* [Computer program]. Chicago, IL: Scientific Software, Inc.
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425, (2001).
- Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.) (1999). *Grading the nation's report card: Evaluating NAEP and transforming assessment of educational progress*. National Academy of Sciences-National Research Council, Board of testing and assessment, Washington, DC.
- Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, 27(4), 15-29.
- Pitoniak, M. J. (2003). *Standard setting methods for complex licensure examinations*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.
- Popham, W. J. (1978). As always proactive. *Journal of Educational Measurement*, 15, 297-300.

- Porter, A. C., Linn, R. L., & Trimble, C. S. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues and Practice*, 24, 32-39.
- Plake, B. S., Huff, K., & Reshetar, R. (2009, April). *Evidence-centered assessment design as a foundation for achievement level descriptor development and standard setting*. Paper presented at the annual meeting of the National Council for Measurement in Education. San Diego, CA.
- Plake, B. S., & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of examinees: An issue in judgmental standard setting. *Educational Assessment*, 7, 87-97.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.
- Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Reckase, M. D. (2006a). A conceptual framework for a psychometric theory of standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, 25, 4-18.
- Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice*, 25, 14-17.
- 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. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Reckase, M. D. (2000, June). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts*. Iowa City, IA: ACT, inc.
- Reckase, M. D. (1998, April). *Setting standards to be consistent with an IRT calibration*. Unpublished manuscript.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph No. 17*.

- Schneider, M. C, Eagan, K. L., Siskind, T., Brailsford, A., & Jones, E. (2009, April). *Concurrence of target descriptors and mapped item demands in achievement levels across time*. Paper presented at the annual meeting of the National Council for Measurement in Education. San Diego, CA.
- Schulz, E. M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice*, 25, 4-13.
- Schulz, E. M., Kolen, M. J, & Nicewander, W. J. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement*, 23, 347-362.
- Schulz, E. M., & Mitzel, H. C. (2005, April). *The Mapmark standard setting method*. Paper presented at the meeting of the National Council for Measurement in Education. Montreal.
- Schulz, E. M., & Mitzel, H. C. (in press). A mapmark method of standard setting as implemented for the National Assessment Governing Board. In E. V. Smith, Jr., and G. E. Stone (Eds.), *Applications of Rasch measurement in criterion-referenced testing*, JAM Press.
- Shen, L. (2001, April). *A comparison of Angoff and Rasch model based item map methods in standard setting*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.
- Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting standards for student achievement*. Stanford, CA: National academy of Education.
- Shepard, L. A. (1994, October). Implications for standard setting of the National Academy evaluation of the National Assessment of Educational Progress achievement levels. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessment*. Washington, DC: National Assessment Governing Board and National Center for Educational Statistics.
- Skaggs, G., Hein, S., & Awuor, R. (2007). Setting passing scores on passage-based tests: A comparison of traditional and single-passage bookmark methods. *Applied Measurement in Education*, 20, 405-426.
- Skaggs, G., & Tessema, A. (2001, April). *Item disorderliness with the Bookmark standard setting procedure*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Sireci, S. G., Hambleton, R. K., & Pitoniak, M. J. (2004). Setting passing scores on licensure exams using direct consensus. *CLEAR Exam Review*, 15, 21-25.

- van der Linden, W. J. (1995). A conceptual analysis of standard-setting in large scale assessments. In *Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Educational Statistics (NCES), Volume II* (pp.97-118). Washington, DC: U. S. Government Printing Office.
- van der Linden, W. J., (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement*, 19, 295-308.
- van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- Verhelst, N., & Kaftandjieva, F. (1999). *A rational method to determine cutoff scores (Research Report 99-07)*. Enschede, The Netherlands: University of Twente, Faculty of Educational Science and Technology, Department of Educational Measurement and Data Analysis.
- Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, 40, 231-253.
- Williams, N. J., & Schulz, E. M. (2005, April). *An investigation of response probability (RP) values used in standard setting*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.
- Wright, B. D. & Masters, G. N. (1981). *Rating scale analysis*. Chicago: MESA press.
- Wright, B. D., & Stone, M. (1979). *Best test design*. Chicago: MESA press.
- Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, 20, 15-25.

MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 03063 1687