You only live up to the standards you set: An evaluation of different approaches to standard setting

Scott N. Strickman
James Madison University

Follow this and additional works at: http://commons.lib.jmu.edu/diss201019

Part of the Educational Assessment, Evaluation, and Research Commons, Educational Methods Commons, Educational Psychology Commons, Higher Education Commons, Quantitative Psychology Commons, and the Social Statistics Commons

Recommended Citation
Strickman, Scott N., "You only live up to the standards you set: An evaluation of different approaches to standard setting" (2017). Dissertations. 151.
http://commons.lib.jmu.edu/diss201019/151

This Dissertation is brought to you for free and open access by the The Graduate School at JMU Scholarly Commons. It has been accepted for inclusion in Dissertations by an authorized administrator of JMU Scholarly Commons. For more information, please contact dc_admin@jmu.edu.
You Only Live Up to the Standards You Set: An Evaluation of Different Approaches to Standard Setting

Scott N. Strickman

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Assessment and Measurement

May 2017

FACULTY COMMITTEE:

Committee Chair: Dena A. Pastor, Ph.D.

Committee Members/Readers:

Keston H. Fulcher, Ph.D.

Allison J. Boykin, Ph.D.
Acknowledgements

The support of so many people made this dissertation possible. First, I’d like to salute my advisor and fellow Texan, Dr. Dena Pastor. I am so grateful for your willingness to confront a dissertation topic that may have been outside of our comfort zones, initially. Your grace, poise, positivity, Texas-sized laugh, and, maybe most importantly, your willingness to “hustle” are unparalleled. You helped me feel unafraid to make mistakes and there’s no way I could have accomplished anything during my time in the program without your support. I always looked forward to our weekly meetings and I truly consider you a friend as well as a mentor.

I also want to thank the other members of my dissertation committee, Dr. Keston Fulcher and Dr. Allison Boykin. Keston, without you, my sentences would consist of far too many prepositions and would be longer than 25 words. Seriously though, I am amazed and inspired by your ability to connect people. You are a fantastic person and a great example of what students should strive to become, both personally and professionally. You genuinely care. Allison, thank you for your encouragement and support in my coursework, research, assistantship responsibilities, and this dissertation. Above all else, I appreciate the variety of topics we could cover in a single conversation—from the nuances of Bayesian data analysis, to stories about your adorable bulldogs, back to bootstrapping reliability techniques.

I would also like to thank all the students, faculty, and staff in CARS. We have been through a lot together. I hope you all feel a sense of pride in helping me become a better student, researcher, and person because each of you has influenced me. Specifically, I want to thank Courtney, Kristen, Thai, and Aaron. You are great friends.
who have always been there for me. I look forward to building even stronger friendships with y’all and can’t wait to see all the great things each of you will accomplish.

Finally, I want to thank my family and friends. To my parents, thank you for teaching me the value of perseverance, dedication, and compassion. Dad, you showed me what it means to work hard in pursuit of my dreams. Mom, you always believed in me, but you reminded me the most important thing I can do is care about people. You also fed most everyone in the program at least once with your frequent “care” packages. To my brother, Tim, and my sister, Steph, I am so proud of the young adults you have become and can’t wait to spend more time with you in the future. To my little puppy, Chloe, thanks for reminding me to to play with stuffed animals and take naps. Your perspective is worth more than you will ever know. Lastly, to my fiancée, Amy: Your support has pushed me farther than I ever dreamed possible. You are my biggest fan, and for that I am forever grateful. You are the best person I know and I am so excited for our future together.
Table of Contents

Acknowledgements........................................................................................................... ii
List of Tables .................................................................................................................... vii
List of Figures ................................................................................................................... viii
Abstract .......................................................................................................................... ix
CHAPTER ONE: Introduction ......................................................................................... 1
  Need for the Current Study ......................................................................................... 5
CHAPTER TWO: Literature Review ............................................................................... 12
  Classification of Standard Setting Methods ............................................................... 12
    Traditional Approaches to Standard Setting ............................................................ 13
      Test-centered .......................................................................................................... 13
        Angoff .................................................................................................................. 14
        Bookmark ............................................................................................................ 17
      Examinee-centered ................................................................................................. 19
        Contrasting Groups ............................................................................................. 19
        Borderline Group ............................................................................................... 21
  Summary of Traditional Approaches to Standard Setting ........................................... 22
  Data-driven Approaches to Standard Setting ............................................................. 23
    Cluster Analysis ........................................................................................................ 24
    Latent Class Analysis ............................................................................................... 27
    Mixture Rasch Model ............................................................................................... 30
  Summary of Data-driven Approaches to Standard Setting .......................................... 34
  Advantages and Disadvantages of Data-driven Approaches ...................................... 35
  Integration of Traditional and Data-driven Approaches to Standard Setting? .......... 39
  Integrated Approaches to Standard Setting ............................................................... 41
    Augmented LCA ..................................................................................................... 41
    Adapted MRM ........................................................................................................ 45
  Summary of Integrated Approaches to Standard Setting ........................................... 46
  Purpose of the Current Study ....................................................................................... 49
  Research Questions ..................................................................................................... 51
CHAPTER THREE: Methods .......................................................................................... 53
Measures ..............................................................................................................................53
   Ethical Reasoning Identification Test-XA (ERIT-XA) .....................................................53
Traditional Standard Setting ...............................................................................................56
   Modified Angoff .................................................................................................................56
Procedures and Participants ...............................................................................................60
Data-driven and Integrated Standard Setting Approaches .............................................62
   General LCA ..................................................................................................................63
   Ordinal LCA ....................................................................................................................63
   Angoff LCA ....................................................................................................................64
Estimation ..........................................................................................................................65
Model Fit ............................................................................................................................67
Classification Accuracy .....................................................................................................68
Analyses Pertaining to Research Questions ..................................................................69
   Research Questions, Set 1 ...............................................................................................69
   Research Questions, Set 2 ...............................................................................................70
   Research Questions, Set 3...............................................................................................70
   Research Questions, Set 4 ...............................................................................................70
   Research Questions, Set 5 ...............................................................................................71
Validity ...............................................................................................................................72
   Auxiliary Variables .........................................................................................................72
      Survey of Ethical Reasoning (SER) ..............................................................................72
      “Dosage” Data ..............................................................................................................74
      Student Opinion Scale (SOS) ......................................................................................74
      SAT Reading Test .........................................................................................................75
   Validity Analyses .............................................................................................................75
CHAPTER FOUR: Results ....................................................................................................77
   Modified Angoff ..............................................................................................................77
   General LCA ..................................................................................................................79
   Ordinal LCA ....................................................................................................................80
   Angoff LCA ....................................................................................................................81
   Research Questions ........................................................................................................82
List of Tables

Table 1. Latent Class Model Comparisons ................................................................. 116
Table 2. Brief Description of 8KQ Provided to Students who take the ERIT-XA .......... 117
Table 3. ERIT-XA Performance Category Descriptions ...................................... 118
Table 4. Student Status, Testing Context, and Sample Size for ERIT-XA data .......... 119
Table 5. ERIT-XA Performance Standards by Round and Rater .......................... 120
Table 6. Median Ratings by Borderline Group for ERIT-XA Items (Round 1) ....... 121
Table 7. Median Ratings by Borderline Group for ERIT-XA Items (Round 2) ........ 123
Table 8. Fit Indices and Entropy for General LCA, Ordinal LCA, and Factor Models .. 127
Table 9. Descriptive Statistics for Modified Angoff and 4-class Ordinal LCA
   Classifications ........................................................................................................ 128
Table 10. Average Posterior Probabilities of Class Membership by Latent Class of the
   Championed Ordinal LCA Solution ......................................................................... 129
Table 11. Comparison of Student Classifications using Modified Angoff and 4-class
   Ordinal LCA ........................................................................................................... 130
Table 12. Summary of Descriptive Statistics for Auxiliary Variables .................... 131
Table 13. Validity Analysis Results for Modified Angoff and 4-class Ordinal LCA .... 132
Table 14. Classifications by Participation in MC Intervention for Modified Angoff ..... 134
Table 15. Classifications by Participation in MC Intervention for 4-Class Ordinal LCA
...................................................................................................................................... 135
List of Figures

Figure 1. Cluster profiles from GED writing skills test.................................136
Figure 2. Cluster profiles from Grade 7 mathematics test............................137
Figure 3. Two-class LCA profile plot of responses to a middle school math test .......138
Figure 4. Five-class MRM profile plot from reading subscale of a language proficiency test..............................................................................................................139
Figure 5. Four-class example plot of non-parallel profile ................................140
Figure 6. Nine steps of a standard setting.......................................................141
Figure 7. Standard setting agenda for the ERIT-XA........................................142
Figure 8. Visual portrayal of “borderline” groups and performance categories for the ERIT-XA standard setting .................................................................143
Figure 9. Survey platform used by SMEs to provide item ratings for each borderline group ............................................................................................................144
Figure 10. Median item ratings provided by SMEs for each borderline group in Rounds 1 and 2 .................................................................................................145
Figure 11. Raw score distribution for subsample of students who took the ERIT-XA ...146
Figure 12. Percent-correct score distribution for subsample of students who took the ERIT-XA ........................................................................................................147
Figure 13. Probability of correct response by item for the ERIT-XA for the 4-class general LCA solution ..................................................................................148
Figure 14. Probability of correct response by item for the ERIT-XA for the 4-class ordinal LCA solution ..................................................................................149
Figure 15. Path diagrams showing the connection between LCA and a factor model ....150
Figure 16. Probability of correct response by item for the ERIT-XA for the one-factor model ........................................................................................................151
Abstract

Interpretation of performance in reference to a standard can provide nuanced, finely-tuned information regarding examinee abilities beyond that of just a total score. However, there is a multitude of ways to set performance standards yet little guidance regarding which method operates best and under what circumstances. Traditional methods are the most common approach adopted in practice and heavily involve subject matter experts (SMEs). Two other approaches have been suggested in the literature as alternative ways to set performance standards, although they have yet to be implemented in practice. Data-driven approaches do not involve SMEs but rather rely solely upon statistical techniques to classify examinees into groups. Integrated approaches are a newer standard setting method that combines judgments provided by SMEs with statistical techniques to inform the creation of performance standards. The primary purpose of this dissertation was to describe and illustrate the traditional, data-driven, and integrated approaches used to establish performance standards on tests. A traditional standard setting was conducted using the modified Angoff procedure. Latent class analysis (LCA)—a data-driven classification technique—was performed in which model parameters were first freely estimated to assess the fit of various general LCA models and later constrained to create ordered groups for various ordinal LCA models. The traditional and data-driven standard setting methods were combined to form an “integrated” approach. SMEs’ ratings of expected examinee performance (derived from the modified Angoff standard setting) were used as item difficulty constraints in an integrated LCA model, the Angoff LCA. The results were used to compare examinee classifications from all three approaches and model-data fit amongst the statistically-
oriented methods. Although classifications were planned for comparison across all three approaches, issues were encountered with the Angoff LCA. Therefore, the comparisons of primary interest were between the modified Angoff and championed LCA model. The results did not offer a clear-cut decision about which approach to champion. Ultimately, the modified Angoff was selected as the most appropriate standard setting approach for the test administered. Important considerations are offered for researchers who wish to use data-driven models to set standards and ideas are proposed for future research.
CHAPTER ONE

Introduction

When tests are administered, the desire is often to order performance along a continuum; to evaluate how test takers fare relative to others. Another purpose of testing is to compare examinee performance to a criterion. The intent of standard setting is to classify examinee scores into ordered proficiency categories in reference to a particular benchmark, or “standard of performance” (Cizek, 1996a; Hambleton, Pitoniak, & Coppella, 2012; Sireci, Robin, & Patelis, 1999). Interpretation of performance in reference to a standard can provide nuanced, finely-tuned information regarding examinee abilities beyond that of just a total score. For instance, performance on a test can be classified into ordered groups—such as Developing, Proficient, and Advanced—to better understand the levels at which examinees are performing.

In a typical standard setting, subject matter experts (SMEs) begin by systematically reviewing the items of an assessment and the characteristics of the examinee population taking the test. To establish cut scores that distinguish levels of performance on a test, SMEs then consider the number of latent (or unobserved) groups that exist in the examinee population, the labels assigned to each group, and the evaluative criteria used to characterize each level of performance. After a standard setting procedure is conducted, it may be determined a score of 80% correct is considered necessary to deem a student Proficient. If the average score is 60% correct, the typical student would not be considered Proficient, but Developing in their knowledge, skills, and abilities measured by the test. Thus, additional meaning can be added to assessment results by establishing standards and labels (e.g., Developing, Proficient, Advanced) to
classify scores into performance categories. However, arbitrary decisions should not be made when setting performance standards if they are to be useful and add value beyond a simple percent-correct or total score. Cizek (1993) described standard setting as “the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance” (p. 100). Clearly, standard setting is much more rigorous than casualty selecting a test score to differentiate examinee performance. Rather, a constellation of factors that affect examinee performance on a test should be taken into consideration in a standard setting study and ultimately used to guide decisions regarding the standards that reflect various levels of performance on a test (Cizek, 1996a).

Three different approaches have generally been used to obtain performance standards. The *Traditional* approaches to standard setting are by far the most common and consist of many methods that can be classified into two general subgroups: (1) Test-centered methods, and (2) Examinee-centered methods. Test-centered methods involve SMEs as raters of item difficulty for a hypothetical group of “borderline” examinees who do not clearly belong to one performance category. Examples of frequently-used test-centered methods include the Angoff (Angoff, 1971) and bookmark (Mitzel, Lewis, Patz, & Green, 2001) procedures, each of which are described in greater detail in Chapter Two. Examinee-centered methods emphasize the familiarity SMEs have with actual groups of students. SMEs draw from their experiences with examinees (e.g., through coursework or as an academic advisor) to classify them into performance groups. Examples of examinee-centered methods that have often been used include the contrasting groups (Livingston & Zieky, 1982) and borderline group (Zieky & Livingston, 1977) procedures,
which are also described further in Chapter Two. A defining feature of the traditional approaches to standard setting is a reliance on SMEs to generate ratings of expected performance that take multiple factors into account. Specifically, SMEs consider performance category descriptions, test content and difficulty, and characteristics of either hypothetical or known examinees to derive cut scores that are used to classify examinees into groups (Zieky, 2012).

Although the traditional approaches to standard setting are popular, there are criticisms associated with their use. Buzzwords like "subjective,” “judgmental,” and “arbitrary” persist in the dialogue regarding the traditional standard setting methods (Cizek, 2012; Hambleton, 1978; Kane, 2001a; Popham, 1978). There is a general level of concern about the inherent subjectivity when cut scores are based primarily on human judgment (i.e., that of SMEs). For instance, rating the difficulty of items for a hypothetical group of “borderline” examinees can be a cognitively complex task (Angoff, 1988; Cizek, 1996a; Fitzpatrick, 1989; Sireci et al., 1999). Similarly, it may be difficult for SMEs to detach from personal feelings to provide an objective assessment of how an examinee will perform on a test (Jaeger, 1989). Thus, the “arbitrary” and “subjective” nature of traditional standard setting methods can be considered a threat to the validity of performance standards (Cizek, 1996). Further, traditional standard settings can be time-intensive and costly, making them particularly difficult to implement. A comprehensive analysis of the challenges associated with traditional standard-setting methods is presented in Chapter Two.

Data-driven approaches to standard setting use statistical techniques to classify examinees into groups based on their responses to the items of a test. Data-driven
approaches to classification have been proposed as an alternative to the traditional standard setting methods for many years, but have yet to be used in practice. The removal of SMEs from the classification process may seem to make the data-driven approaches more objective than traditional methods. However, subjectivity is not eliminated entirely with data-driven approaches (Sireci et al., 1999). Data-driven classification techniques will always partition examinees into groups, regardless of whether they truly exist. It is ultimately up to the researcher’s discretion to determine which groups represent meaningful classification of examinees (Sireci et al., 1999). Further, the groups suggested by data-analytic techniques may differ in number and nature from the performance groups put forth by policymakers and other stakeholders. Five groups may emerge from the results of a statistical technique even though only three performance categories were specified by policymakers. Or, the results of data-driven classification techniques may indicate different groups of examinees are not ordered along a continuum, whereas the performance categories established by policymakers are assumed to be ordered from low to high. Three statistical techniques have been investigated as potential data-driven methods for standard setting, and their proposed use is detailed in Chapter Two.

The Integrated approaches to standard setting combine information from SMEs with statistical techniques to classify examinees into performance groups. Examinees are classified into groups on the basis of SMEs’ judgment and on their response pattern across items of a test. The integration of both expert judgment from a traditional standard setting approach and statistical results from a data-driven approach offers multiple sources of information when making classifications of examinees. The combination of subjective information (from SMEs) and empirical information (from data) helps
integrated approaches overcome some of the issues—such as “arbitrariness,” “subjectivity,” or “meaningfulness”—associated with other standard setting methods (Templin & Jiao, 2012). That is, researchers may make more informed and improved classification decisions using an integrated approach to standard setting than with another method that relies exclusively on either SMEs or data-driven techniques.

Two integrated methods have been proposed that incorporate SMEs’ judgments from a traditional standard setting into a statistical technique, specifically, a mixture model. Mixture models are used to identify latent classes (or unobserved groups) of individuals that exist in a larger population of examinees. One integrated approach uses SMEs’ classifications of test takers from a traditional, examinee-centered standard setting to inform a latent class analysis (LCA) model, which is a type of mixture model (Templin, Poggio, Irwin, & Henson, 2007). Another integrated approach operates under a similar principle but incorporates the same classifications from SMEs into a mixture Rasch model (MRM; Templin, Cohen, & Henson, 2008). Although neither integrated method has been used in practice to set performance standards for a test, both offer promising alternatives to standard setting in higher education. Further consideration of the integrated approaches is offered in Chapter Two.

Need for the Current Study

Now that we have a multitude of standard setting approaches to choose from, the obvious question is which should be used and when? The present study attempts to answer important questions that emerge from use of the different standard setting approaches. For example, do the results from one standard setting approach approximate those found using another approach? How does the number of performance categories
established for a traditional standard setting compare to those derived using statistical techniques? Does the integration of multiple sources of information lead to different conclusions than using a single source? Are the various approaches to standard setting differentially related to variables that could be used as external validation of performance standards?

The purpose of this study is to apply each of the standard setting approaches to a 50-item, multiple-choice examination of ethical reasoning used in higher education. The results from a traditional approach, two data-driven approaches, and an integrated approach to standard setting were evaluated to answer some of the above questions. The modified Angoff procedure, a test-centered method, was implemented for the traditional standard setting. SMEs considered a hypothetical pool of examinees (from a larger population of students at a four-year institution in the Mid-Atlantic) on the border of either the Developing/Proficient or the Proficient/Advanced performance categories in specific components of ethical reasoning knowledge and abilities. SMEs provided estimates of expected performance for the two groups of borderline examinees and their ratings were subsequently used to establish two performance standards on the test.

Two data-driven standard setting approaches were also considered: a general LCA and ordinal LCA. The two approaches both used a LCA model to classify examinees into performance groups based on responses to all items of the ethical reasoning test. The general LCA model aligns with the manner in which latent class analysis has been used as a data-driven approach in the literature. The general LCA model allows the number and nature of groupings of examinees to take any form; that is, the number of groups resulting from the general LCA does not need to be specified in advance and groups may
be ordered or unordered. Because LCA is capable of yielding unordered groups—which does not correspond to the performance category descriptions created for the ethical reasoning test—another LCA model that constrained groups to be ordinal was considered. The ordinal LCA forces groups to be ordered but does not limit the number of groups in the model; that is, models with different numbers of performance groups can be investigated with those groups forced to be ordered.

A new integrated standard setting approach was evaluated as well. The integrated approaches proposed in the literature thus far have only used SME judgments obtained from traditional, examinee-centered methods to standard setting. The integrated method in the current study used SME judgments obtained from the modified Angoff procedure, which is a traditional, test-centered standard setting method. Specifically, the judgments provided by SMEs regarding the performance of borderline examinees on each item—which were collected during the modified Angoff procedure described earlier—were used to inform the item difficulties for latent classes in a LCA. Unlike the two data-driven LCAs tested in this study, the Angoff LCA requires the number of latent classes to be equal to the number of performance categories specified for the traditional standard setting (i.e., the modified Angoff). The number of groups must match between the two approaches because the item difficulties for each latent class in the Angoff LCA are constrained to align with the ratings provided by SMEs from the modified Angoff procedure. Because three performance groups were specified for the modified Angoff, SMEs provided ratings of expected performance for two groups of borderline examinees. These ratings were used to impose constraints on the class-specific item difficulties in the Angoff LCA. Specifically, the item difficulty for each item in the first class was
constrained to be lower than the item difficulty provided by SMEs for the 

*Developing/Proficient* borderline group. Additionally, the item difficulty for each item in the second class of the Angoff LCA was constrained to be higher than the item difficulty provided by SMEs for the *Developing/Proficient* borderline group, but lower than the SME rating for the *Proficient/Advanced* borderline group. Finally, the item difficulty in the third class was constrained to be higher than the SME rating for the 

*Proficient/Advanced* borderline group. As a result of these constraints, the estimated item difficulties for each class are ordered from low (in the first class) to high (in the third class) and in a way that aligns with SMEs’ expectations of performance.

There has yet to be a study that compares the results from a traditional, data-driven, and integrated standard setting. A comparison of particular interest in this study is how the data-driven and integrated approaches fare relative to one another, especially considering they have yet to be used in practice. The data-driven and integrated approaches are each perceived to minimize the amount of subjectivity involved in the classification of examinees, compared to a traditional method. If the element of subjectivity in standard setting is reduced by using a data-driven or integrated approach, it may be a worthwhile endeavor to pursue in some situations.

One way the data-driven and integrated standard setting approaches are compared in the present study is by analyzing how the models differ in their fit to the data. Although the number of parameters estimated does not vary across models when the same number of classes is specified, it was suspected that the differing constraints of the data-driven and integrated approaches would impact model-data fit. The least constrained model (i.e., the general LCA) was expected to fit better than the more constrained ordinal
LCA and Angoff LCA models. The general LCA constitutes the least constrained model in the current study because its parameters (e.g., proportion of examinees in each group and item difficulties for each group) are allowed to freely vary. There are no restrictions on the nature of the groups formed in a general LCA; they may be ordered or unordered. Conversely, the ordinal LCA and Angoff LCA models both have constraints imposed that force groups to be ordered. Although these models yield results that better align with the performance category descriptions than the general LCA, the data will always be best characterized by a model with fewer constraints. In this case, the general LCA has the least constraints and will fit the data best. A question of interest in the current study is whether the general LCA, ordinal LCA, and Angoff LCA show similar fit to the data. In other words, will the ordinal and Angoff LCA models fit the data well enough to be worth the added constraints?

An essential part of the standard setting process is examination of the validity of the decisions made to separate examinees into different groups. Validity refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA, APA, NCME, p. 9). The appropriateness of performance groups and the cut scores used to distinguish them is the aim of validation in standard setting (Kane, 2001a). Further analyses beyond the standard setting are often conducted to compare whether the classifications based on cut scores are consistent with other indicators of an examinee’s performance ability (Kane, 2001a). If findings show examinees classified into distinct performance categories have different outcomes on other theoretically related constructs, there is additional support for the classifications resulting from the standard setting method. For example, examinees in different
performance groups may deviate in their attitudes toward ethical reasoning, the relative importance they assign to ethical reasoning in their lives, or their amount of exposure to the ethical reasoning curriculum. Although none of these findings are likely to be definitive, a consistent body of evidence may be helpful when making a decision to advocate for one method or approach over another (Kane, 2001a). Thus, in addition to analyzing how the models from the data-driven and integrated approaches compare in terms of fit to the data, the relation between performance group membership and other external variables is also investigated within and across standard setting methods as part of the current study. Research questions pertaining to the relations between group membership and other variables examined in the current study include: How do the groupings from the various standard setting approaches relate to other variables? Does one approach yield relations with other variables that are more consistent with expectations than another?

In summary, the purpose of the current study is to describe and illustrate the different approaches to standard setting. The three different approaches have unique qualities and a study assessing the results from each could have implications for setting performance standards. Additionally, a novel integrated approach—the first to use SMEs’ judgments from a traditional, test-centered standard setting method—is explored as a viable option for classifying examinees into performance groups. Important questions about standard setting are addressed in the current study and conclusions are drawn about the various approaches. For instance, how similar are the results from different standard setting approaches used on the same test? Are there particular conditions under which stakeholders may want to choose one standard setting approach over another? If so, how
and when should each of the approaches be applied to standard setting? This paper provides a comprehensive review of the approaches to standard setting that utilize statistical models and offers insight into the many considerations that should be contemplated when selecting a particular method or approach to set performance standards on a test.
CHAPTER TWO

Literature Review

The literature review that follows will begin with a brief discussion of the history of standard setting method classifications. Attention will then turn to the existing approaches used for standard setting. The specifics of commonly-used standard setting methods within each approach will be explored, including logistical details of implementation, how cut scores are generated to establish different performance categories, and research investigating the use of popular methods. A summary will be offered at the conclusion of each section to encapsulate the rationale of each approach and the advantages/disadvantages associated with their use. Finally, the prospect of an alternative approach to standard setting will be proposed that integrates current approaches, with examples from recent practice serving as a springboard to the purpose and associated research questions of this study.

Classification of Standard Setting Methods

The conversation about general types of standard setting methods has long been discussed in binary terms. Jaeger (1989) proposed a classification scheme of standard setting methods that has received support from key figures in the field (Cizek, 1996a; Kane, 1994b). Test-centered methods involve judgments made by subject matter experts (SMEs) regarding the items on a test, whereas examinee-centered methods focus on SMEs’ judgments about the abilities of test takers. Test-centered and examinee-centered methods are referred to in this paper as Traditional standard setting approaches because they rely on judgments from SMEs to derive a cut score, which has been a staple of standard setting since its origin. Data-driven approaches rely on statistical techniques to
classify examinees into groups based on their responses to the items of a test. Studies have used statistical methods, such as cluster analysis (Sireci, 1995; Sireci, 2001; Sireci, Robin, & Patelis, 1999), latent class analysis (Brown, 2007), and mixture Rasch models (Jiao, Lissitz, Macready, Wang, & Liang, 2011), to provide demonstrations of how data-driven techniques could be used to create performance standards; however, these methods have not been used to set standards in practice. A recent development in standard setting is an Integrated approach, which combines judgments made by SMEs and statistical classification techniques to establish cut scores informed by both traditional and data-driven approaches. These three categories (i.e., Traditional, Data-driven, and Integrated) encompass the majority of contemporary standard setting approaches. The purpose of the literature review is to describe these three types of standard setting approaches, with particular consideration given to the benefits/challenges of each approach—specifically, in relation to one another.

**Traditional Approaches to Standard Setting**

Traditional approaches are based on judgments provided by SMEs and are used in a large majority of standard settings. The traditional approaches will be broken down into test-centered methods, focusing on the modified Angoff and bookmark procedures as examples, and examinee-centered methods, focusing on the contrasting groups and borderline group procedures as examples.

**Test-centered.** Test-centered approaches to standard setting involve SMEs who make ratings regarding the difficulty of each item on a test for a hypothetical group of examinees. SMEs indicate how difficult they perceive each item to be for borderline examinees taking the test (Sireci et al., 1999). A borderline examinee is a test taker who
is on the cusp of being in either of any two adjacent performance categories, such as *Developing/Proficient* or *Proficient/Advanced*. Two of the most commonly used test-centered techniques are the *Angoff* method and the *bookmark* method. The details of each method are described in the next two sections.

**Angoff.** The Angoff standard setting method (Angoff, 1971) is one of the most popular procedures for a test comprised predominantly of multiple-choice items (Hambleton & Pitoniak, 2006; Kane, 1994b; Mehrens, 1994). In the Angoff procedure, SMEs are asked to independently rate the probability borderline examinees will correctly answer the items of an assessment. For example, imagine faculty at an institution of higher education desire to categorize students into three performance groups—such as *Developing/Proficient/Advanced*—for a 50-item multiple-choice test. If the Angoff method were used, SMEs advance through the test item-by-item and assign likelihood ratings (i.e., probabilities) that a borderline student will correctly answer that item. The “minimally proficient” students are those on the border of *Developing/Proficient*, and “minimally advanced” students are on the border of *Proficient/Advanced*.

Because students will be classified into three groups, two cut scores are needed to separate performance into the desired categories. SMEs are first gathered to discuss the knowledge, skills, and abilities (KSAs) that should be considered characteristic of each group, with this discussion informed by the descriptors (or evaluative criteria) of each performance category. The performance category descriptions might be created by a panel before a standard setting or as the first part of a standard setting (Hambleton & Pitoniak, 2006). Regardless, it is important for SMEs to be trained on the descriptors (or evaluative criteria) of each performance category before item ratings begin so all SMEs
are prepared to make knowledgeable decisions about group membership (Hambleton & Pitoniak, 2006).

After SMEs have been trained regarding the relevant KSAs for each performance group, they go through the items while keeping in mind the qualifications of a *minimally proficient* student (i.e., one on the border of Developing/Proficient) taking the test. SMEs independently assign a probability of correct response from a minimally proficient student to each of the 50 items (e.g., 0.50 probability of correct response to item 1, 0.75 probability of correct response to item 2, and so on). The item probabilities are summed for each SME, and either the mean or median across all SMEs is used to create a cut score separating the Developing/Proficient performance categories. Students below the cut score are classified as Developing, but those above the cut score are not yet classified because they might be in either the Proficient or Advanced group. Thus, a second cut score is needed to separate the Proficient/Advanced performance groups in this example, so SMEs must also assign likelihood ratings of a correct response to each item for another borderline group—*minimally advanced* students taking the test. SMEs begin the process of rating item difficulties again, providing a probability of correct response for a minimally advanced student to the first item, the second item, etc. The item probabilities are summed for each SME, and either the mean or median across all SMEs is used to create a cut score distinguishing the Proficient/Advanced performance categories. Students at or exceeding the cut score are placed into the Advanced group, and students below the cut score are classified as Proficient. The result of SMEs’ work is the classification of students into three ordered groups.
A modified version of the Angoff method involves multiple rounds of SMEs independently estimating, for each item, the proportion of borderline examinees who will answer correctly. Each round of rating is followed by discussion amongst the SMEs in which they confer about the rationale behind their ratings. Using the modified Angoff with the earlier example, multiple rounds of discussion would occur for each of the individual item probabilities and resulting cut scores. Additionally, impact data are often presented by facilitators of the standard setting. The impact data can show SMEs practical consequences of the standards they have just established, such as the percentage of examinees classified into each performance category. After feedback and data are shared, SMEs are given the opportunity to adjust their ratings. The opportunity to provide feedback between rounds is a useful way of reducing variation in item probability ratings, which has been shown to increase reliability and “promotes reasonable conceptualizations of anticipated examinee performance” (Cizek, 1996b, p. 23). Research findings have indicated the Angoff method is simple to explain (Mills & Melican, 1988) and offers the best balance between technical and practical application (Berk, 1986).

A common critique of the Angoff procedure is the difficulty of conceptualizing a borderline examinee (e.g., Brandon, 2004; Hambleton & Pitoniak, 2006; Plake & Impara, 2001). The National Academy of Education (1993) called the process of estimating how borderline test takers will fare on test items a “nearly impossible task” (p. xxiv). Further complicating matters is the cognitively demanding task of assigning the likelihood that “borderline” examinees (e.g., minimally proficient students) will correctly answer specific questions (Angoff, 1988; Cizek, 1996a; Fitzpatrick, 1989; Sireci et al., 1999). Consistency of standards across panels and samples of examinees is paramount to the
validity of interpretations made from performance standards. Subjective interpretations of what defines, and is expected of, a borderline examinee threatens the validity of cut scores. Additionally, if a test has many items, rater fatigue may become an issue. Recall the modified Angoff example outlined earlier in this section, in which SMEs set two cut scores on a test comprised of 50 items. Not only do those SMEs have to independently rate probabilities for 50 items twice (once for \textit{minimally proficient} students, and again for \textit{minimally advanced} students), but they must also discuss their item ratings with the group afterward. And then, SMEs would be asked to go back through all the items again and amend any ratings they would like to change, with the potential of further rounds if adequate consistency is not reached! This certainly seems like a draining and protracted task for all participants involved in the standard setting.

\textit{Bookmark.} The bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001) has also been a popular standard setting method (Karantonis & Sireci, 2006; Zieky, 2012). In the bookmark method, items are presented in order of increasing difficulty (usually based on item response theory [IRT] item-difficulty estimates) in what is known as the Ordered Item Booklet (OIB). SMEs place a “bookmark” between two items, such that they expect a borderline student to correctly answer all items preceding the bookmark based on a predetermined Response Probability (RP). RPs reflect the probability with which a borderline student (or group of students) is expected to correctly answer an item. RPs remain the same for all items of a test and are often established as either \(0.50\) or \(0.67\)—which equates to a 50\% or 67\% likelihood of success on an item (Wyse, 2015). Returning to the earlier example involving three performance categories (i.e., \textit{Developing/Proficient/Advanced}), consider a bookmark standard setting with a RP of
0.67. Just as with the modified Angoff, two cut scores will be generated because we have three performance categories and, thus, two groups of borderline students. First, the SMEs independently consider whether a minimally proficient student has a 0.67 probability or greater of obtaining the correct response for the least difficult item of a test (based on IRT estimates). If the probability of a correct response to that item is greater than or equal to 0.67, the next item in the OIB is evaluated. SMEs continue through the OIB until they encounter an item they believe a minimally proficient student will correctly answer less than 67% of the time, at which point a “bookmark” is placed. Individual performance standards for each SME are generated by estimating the theta value associated with the 0.67 probability of a correct response to the item preceding each SME’s bookmark, and the average or median theta value of all SMEs is used to establish a cut score. The theta cut score can be transformed to another scale, if desired, to aid interpretation of results for stakeholders. Because a second cut score needs to be established in our example to distinguish between Proficient and Advanced students, the process of reviewing items continues until each SME identifies a second bookmark, indicating a less than 67% chance of correct response on that item (and thereby, subsequent items as well) for a minimally advanced student.

The use of RP’s is necessary to transform performance category descriptions into cut scores. Applying different RP’s to the same test and sample, however, is likely to affect the cut score in different ways (Lewis, Mitzel, Mercado, & Schulz, 2012; Williams & Schulz, 2005). Thus, the classification of examinees into performance groups is potentially dependent upon the RP value assigned, and different RP’s are likely to produce inconsistent percentages of examinees in each performance group (Haertel &
Lorie, 2004; Wyse, 2011). As such, claims have arisen about the arbitrary nature of selecting a particular RP over another and the impact of RP selection on setting a performance standard (Karantonis & Sireci, 2006).

**Examinee-centered.** The impetus behind the development of examinee-centered approaches was to replace the difficult cognitive task of judging the performance of borderline students on an array of items—as is done using test-centered methods—with a simpler task of making categorical judgments about the performance level of examinees that are familiar to SMEs (Zieky, 2012). Thus, examinee-centered techniques differ from test-centered methods in that cut scores are derived by having SMEs evaluate the ability level of actual examinees, rather than the probability of correct response from a hypothetical group (Sireci et al., 1999). For all examinee-centered approaches, SMEs rate the performance level of examinees based on prior experience with the group of test takers—such as students who took a course taught by a SME involved in the standard setting. According to Mehrens (1994) and Brandon (2002), the *contrasting groups* method and *borderline group* method are two of the most popular examinee-centered approaches to standard setting in higher education. Each of these methods is discussed in the next two sections.

**Contrasting Groups.** In the contrasting groups method described by Livingston and Zieky (1982), SMEs review performance category descriptors and identify examinees they believe are clearly within a performance group. SMEs are asked to divide examinees into groups based on prior knowledge of the general skills and abilities possessed by that group of examinees. All of the classifications done by SMEs are completed without knowing the examinees actual test scores. The familiarity SMEs have
with the examinees being classified has been said to make the process of judging performance adequacy simpler than test-centered methods (Livingston & Zieky, 1982). Back to our running example, SMEs must evaluate the criteria for Developing, Proficient, and Advanced students and determine which group is most likely for each student. Then, the test score distributions of examinees classified into adjacent groups by SMEs (e.g., Developing/Proficient, Proficient/Advanced) are contrasted to select the performance standards (Hambleton & Pitoniak, 2006). Cut scores are typically calculated by plotting the test score distributions of each group, identifying where adjacent distributions overlap, and calculating the point at which half (50%) of the examinees within the overlap are judged to meet the performance standard (Hambleton & Pitoniak, 2006), though there are other techniques that have been used (Koffler, 1980; Livingston & Zieky, 1989; Webb & Miller, 1995).

A likely reason for the popularity of the contrasting groups method is the ease with which SMEs are able to identify familiar examinees as either above or below a standard, as opposed to the abstract nature of test-centered methods (Sireci et al., 1999). SMEs are able to consider an examinee’s entire body of work when making classification decisions, rather than focusing on the items or results from a single test. This also leads us to a major concern with examinee-centered approaches, in general. SMEs may allow their experiences with particular test takers to bias their perceptions of examinee proficiency. For example, SMEs may consider factors irrelevant to the test content, such as unrelated cognitive or noncognitive abilities, to determine whether a student is Developing, Proficient, or Advanced. Likewise, personal relationships with examinees may sway SMEs in one direction or the other. Cizek (1996a) suggested an extra layer of
validation is needed to authenticate the validity and dependability of the criterion judgments made by SMEs when using the contrasting groups method to set performance standards: “For example, judgments assigning examinees to ‘known’ master or nonmaster groups are fallible. It is equally necessary to examine the adequacy of these classifications as it is to examine the psychometric characteristics of the predictor (i.e., the examination)” (pg. 25).

**Borderline Group.** The borderline group method (Zieky & Livingston, 1977) uses SMEs to identify a group of actual test takers believed to be on the boundary of two performance categories; hence, it is called the “borderline group” method. Just like the contrasting groups method, it is vital SMEs understand the descriptors of each performance level and have intimate knowledge of the population being assessed (Cizek, 1996a). Only students identified to be on the border of adjacent groups, such as Developing and Proficient, are used to generate a cut score that separates those two groups. If more than two performance categories exist, a situation similar to the earlier application of the Angoff method applies. Two borderline groups—those on the edge of Developing/Proficient and those on the edge of Proficient/Advanced—would be considered, the difference being there is no need for an investigation of every item on the test with the borderline group method. The mean or median score of test takers in each borderline group is commonly used to establish performance standards.

As with test-centered approaches, categorization of examinees as “borderline” is not a straightforward task. Even if the person doing the rating is familiar with test takers, the process of simultaneously conceptualizing and identifying students who are not quite “advanced” but seem to be more “proficient” than their peers can be difficult. SMEs may
draw (knowingly or not) from information outside of test content in order to simplify classification of examinees (Cizek, 2012; Jaeger, 1989). Thus, there is a level of subjectivity required, which may result in misclassification of Proficient students as Developing, for example. Another important note is that it may be challenging to find SMEs who are familiar enough with examinees to make judgments about their ability level, or it may be difficult to obtain a large sample of borderline students (Hambleton & Pitoniak, 2006).

**Summary of Traditional Approaches to Standard Setting**

Although there are differences between test-centered and examinee-centered methods, it is important to remember they ultimately fall under the same umbrella of standard setting approaches (i.e., “traditional” approaches that use SMEs to set performance standards). The reliance on human judgment in traditional approaches introduces many challenges. The use of SMEs to make judgments—whether it is about the performance of hypothetical examinees or regarding classification of real students—is often perceived as subjective and can lead to different performance standards. The major issue with test-centered approaches, such as the Angoff or bookmark method, is the difficulty inherent in conceptualizing and defining a population of “borderline” examinees. The potential misclassification of students in examinee-centered approaches is also a real threat to validity. A plethora of texts and research have suggested an instability of cut scores exists across traditional standard setting methods, and even within single methods (e.g., Angoff, 1988; Brandon, 2004; Cizek, 1993; Cizek, 2001; Hambleton & Pitoniak, 2006; Koffler, 1980; Shepard, 1980; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Sireci et al., 1999).
On top of concerns about subjectivity and replicability, there are logistical and time considerations that make traditional approaches resource-intensive. Amongst the tasks needed to conduct a successful standard setting, facilitators must recruit the right (and enough) SMEs, work with SMEs to prepare descriptions of performance categories, train SMEs well on the selected standard setting method, hope SMEs can meet the cognitive demands of the task, and guide SMEs to achieve consistency and agreement. Plus, hefty financial and human resources are required to procure not only SMEs, but also facilitators to lead the standard setting, conduct the training of SMEs, and to reserve a building and pay for food.

In summary, the subjective nature of traditional approaches to standard setting leaves them vulnerable to criticism, rightfully or not. And the potentially costly venture of traditional standard setting approaches makes them difficult to warrant, especially more than once. Thus, an interesting question arises in regard to Jaeger’s (1989) classification scheme of traditional standard setting methods: Should judgments about competency only be formed on the basis of either test content or prior experiences with examinees?

**Data-driven Approaches to Standard Setting**

To reiterate, the goal of standard setting is classification—we want to classify examinees into groups based on their test performance. If our ultimate goal is classification, we may choose to turn to the many numerical and statistical modeling techniques that can be used for classification, which include cluster analysis and various kinds of mixture models (e.g., LCA, latent profile analysis [LPA], factor mixture models, MRM). For instance, the goal of cluster analysis is to classify people into groups based
on their values on a set of variables. Persons in the same group should have similar
values on the set of variables, and persons in different groups should have dissimilar
values on the set of variables (Pastor & Erbacher, in press). Thus, groups are not known
beforehand, but emerge as a result of using the classification technique. Such
classification techniques have been used in a wide range of research areas to identify the
number and nature of groups underlying a set of variables and to classify persons into
groups.

Not surprisingly, in recent years, researchers have proposed the application of
classification techniques to standard setting based solely on data. Data-driven approaches
to standard setting, like classification techniques, rely on numerical and statistical
modeling to classify examinees into groups. Data-driven approaches employ existing test
data as input into the analysis and utilize a numerical or statistical procedure to: 1) find
the number of groups (i.e., performance categories), 2) describe the nature of the groups,
and 3) classify examinees into groups. Three data-driven approaches to standard setting
described in this portion of the literature review are cluster analysis (Sireci, 1995; Sireci,
2001; Sireci, Robin, & Patelis, 1999), LCA (Brown, 2007), and MRM (Jiao et al., 2011).
Demonstrations of how each researcher/set of researchers propose using these data-driven
approaches for standard setting are presented in the next three sections.

**Cluster Analysis.** Cluster analysis is a statistical technique used to group
examinees with similar performance into homogenous clusters. The intent is to group
people with similar values on a set of variables into the same group, and people with
dissimilar values into different groups (Pastor & Erbacher, in press). Cluster analysis is
different from the other data-driven approaches to standard setting because it (usually) is
not based on a statistical model. Rather, cluster analysis is a numerical algorithm that can be used to reduce a large number of observations into smaller groups (Pastor & Erbacher, in press). The ultimate goal of the researcher using cluster analysis is to minimize within-group variability and maximize between-group variability, yielding the most parsimonious yet still meaningful number of clusters that best characterizes the data (Sireci et al., 1999; Sireci, 2001). Many variations of cluster analysis exist, and there are numerous choices the analyst must make to choose a particular procedure. Because a description of the various techniques for cluster analysis is beyond the scope of this study, readers are referenced to Everitt, Landau, Leese, and Stahl (2011) for an overview.

Sireci (1995) appears to be the first researcher to demonstrate how cluster analysis can be applied to standard setting. He performed cluster analysis using two samples of high school seniors who took the GED writing skills test. The test was comprised of 50 multiple-choice items, covering three content areas (Mechanics, Sentence structure, and Usage), as well as a writing sample (Essay). Two performance categories (Pass/Fail) were established for the test before Sireci’s study, with the 30th percentile of the total test score serving as the performance standard separating the Pass/Fail groups. Sireci (1995) applied cluster analysis to the two samples taking the GED writing skills test by using subscale scores (Essay, Mechanics, Sentence Structure, and Usage) as input into the analysis. Although Sireci used the subscale scores as input into the cluster analysis algorithms, it should be noted they are not the sole type of information that can be entered into the analysis; items or orthogonal factor scores can also be used, if desired (Sireci et al., 1999). Results of the analysis revealed five clusters (groups) as the preferred solution in both GED samples. Sireci (1995) cited cluster profile plots as useful
evidence to show SMEs participating in a standard setting. The cluster profile plots for one of the GED samples \((n = 511)\) from the Sireci (1995) study are displayed in Figure 1. Sireci considered the third cluster (C3) in Figure 1 to represent a borderline group of examinees and used the median of their test scores as the cut score. The resulting cut score found by Sireci (32\textsuperscript{nd} percentile) was very similar to the existing standard already in place (30\textsuperscript{th} percentile). Sireci also provided validity evidence by calculating the point-biserial correlation between the Pass/Fail classification based on the cluster analysis and self-reported grades in a high school English composition class \((r_{pb} = 0.38)\). He concluded the results were in alignment with expectations.

Sireci et al. (1999) provided another demonstration of using cluster analysis to set performance standards. Cluster analysis was performed using two samples of 7\textsuperscript{th} grade students who took a statewide Grade 7 reading, writing, and mathematics achievement test. Three performance categories (Intervention, Proficient, and Excellence) were already established for the Grade 4, 6, and 8 versions of the test using the modified Angoff method, and those proficiency groupings were adapted for use with the 7\textsuperscript{th} grade students. Sireci et al. (1999) applied cluster analysis to the two test samples comprised of 7\textsuperscript{th} graders, using only mathematics subscale scores as input into the analysis. Results of the analysis favored a solution with three ordered clusters (groups) in both samples. The cluster profile plots for the Grade 7 mathematics subscales from the Sireci et al. (1999) study are displayed in Figure 2. Sireci et al. (1999) interpreted the groups derived from the cluster analysis to be representative of the three proficiency levels the test was designed to measure (because the cluster profile plots were ordered). As a result, Sireci et
al. (1999) used total test scores for adjacent clusters to set the performance standards (i.e., cut scores).

Sireci et al. (1999) compared the standards derived from the cluster analysis to previously-established performance standards (via the modified Angoff) already in use by the school district. Ultimately, the suggested cut scores and student classifications derived from the cluster-analytic solution (i.e., Intervention/Proficient/Excellence) were similar to those derived by the local school district using the modified Angoff. The cut scores established using the cluster-analytic method classified about 90% of students into the same proficiency grouping as the school district. Sireci et al. (1999) also provided validity evidence by correlating student classification based on the cluster analysis (i.e., Intervention/Proficient/Excellence) with final grades obtained by students in their mathematics courses and total scores on the Grade 7 mathematics achievement test. The Spearman rank-order correlation between student classification and math grades was 0.69 for the cluster-analytic method, and the Pearson correlation between student classification and total mathematics test score was 0.94 for the cluster-analytic method.

**Latent Class Analysis.** Latent class analysis (LCA; Lazarsfeld & Henry, 1968) belongs to a family of statistical models known as mixture models. Mixture models are used when a population is considered to consist of a mixture of unknown groups (i.e., latent classes) that differ in the parameters of the statistical model being used to characterize the data (McLachlan & Peel, 2000). Before going into the details of mixture models, including LCA, it is important to convey how these models function as classification techniques. Like cluster analysis, the number of classes is not known beforehand. Various models, each differing in the number of classes specified, are fit to
the data and statistical indices are combined with judgment by the analyst to decide which solution to champion. Say stakeholders are intrigued by the possibility of either two (Developing/Proficient) or three (Developing/Proficient/Advanced) distinct levels of performance for students taking a test (in current or prior use). In this situation, LCA models differing in their number of classes are fit to the data. Information criteria (e.g., log likelihood, Bayesian Information Criteria, Sample Size Adjusted-Bayesian Information Criteria), and statistical significance tests (e.g., Lo-Mendell-Rubin and Bootstrapping Likelihood Ratio Test) regarding the relative fit of the model are used to evaluate which model fits the data best. Thus, if a 2-class model is championed, the results imply there are two, rather than three, possible performance categories.

Once a solution is championed, the class-specific parameters from the model are used to characterize each group and class weights are used to estimate the proportion of the population in each class. When scored item responses (i.e., correct/incorrect) are used as input into a LCA, the class-specific parameters are item difficulties—the probability of getting a correct response on each item. These class-specific parameters are used to understand similarities and differences across classes, a process facilitated by a profile plot. The posterior probability of membership in each class can also be calculated after model parameters are estimated and used to assign individuals to their most likely group. A general C class mixture model is shown in Equation 1.

$$P(x_i = 1) = \sum_{c=1}^{C} \rho_c P(x_i = 1|c)$$

(1)

The marginal probability of a correct response for item i is equal to the weighted sum of the class-specific or conditional probabilities of a correct response in each class ($P(x_i = 1|c)$), where the weights ($\rho_c$) are the proportion of the population in each class.
Only $C-1$ class weights are estimated because the weights are constrained to be positive and sum to one. Equation 2 shows the probability of a correct response to any item in LCA, where $\tau_{ic}$ represents the item difficulty (on the logit scale) for item $i$ in class $c$ and $\pi_{ic}$ is the probability of a correct response to item $i$ in class $c$.

\[ P(x_i = 1|c) = \pi_{ic} = \frac{\exp(-\tau_{ic})}{1+\exp(-\tau_{ic})} \]  

(2)

To illustrate how LCA functions as a classification technique, consider Brown’s (2007) proposed use of LCA as a data-driven approach to standard setting. In one of his analyses, Brown (2007) used the dichotomous form (i.e., right/wrong) of 12 mathematics items for seventh- and eighth-grade students as input into the LCA. Brown (2007) assessed the fit of a one-class, two-class, and three-class model to the data. Several indexes used by Brown to compare the fit of the three models are presented in Table 1. Strongest support was found for the two-class model, as indicated by higher values on the log likelihood index and lower values on the Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978) measures. The class weights indicated 54% of examinees in the population were in Class 1 and 46% were in Class 2. The profile plot presented for the championed model in Figure 3 shows the item difficulty values are different between the two groups. Although the two classes performed fairly similarly on some items (Items 3 and 8), the general trend indicated items were easier for students in Class 1 compared to students in Class 2. Because the pattern of responses across all items indicates the two classes are ordered, students in Class 1 are believed to have higher mathematics ability than students in Class 2.

Results of Brown’s (2007) LCA were not used to set standards on the total score scale; instead, posterior probabilities of class membership were calculated for each
examinee and used to assign individuals to their most likely group. Brown (2007) compared results from his LCA classifications to two traditional standard setting approaches that used SMEs, including the Angoff method. The standards established using the Angoff method were based on four performance categories. To compare the classifications from the LCA and the Angoff, Brown collapsed the upper two Angoff performance categories into one group and the lower two into a second group. Brown, presumably, did this in order to effectively compare the championed 2-class LCA solution to Angoff results with only two groups (instead of the original four performance categories). A high percentage of agreement was found between LCA and the Angoff method (92.2%) in the classifications of students. Likewise, 77.1% agreement was found between LCA and Jaeger’s profile rating method (1995), the other traditional approach used for comparison. The overall level of agreement between LCA and the traditional methods was 85.7%, indicating the statistical model rendered similar classification decisions as the traditional approaches using SMEs (Brown, 2007). Brown (2007) asserted the similarity in categorization of students between LCA and the traditional methods supports the use of data-driven approaches to standard setting that could at least be used to complement SMEs’ ratings.

**Mixture Rasch Model.** Like LCA, the mixture Rasch model (MRM; Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990) can also be used to analyze test data comprised of multiple latent populations. MRMs are used to fit various numbers of classes to the data, and statistical indices inform the researcher which model to champion—just as is done with LCA. Specifically, the MRM integrates the Rasch measurement model (Rasch, 1960) and a latent class model. The Rasch measurement
model assumes a quantitative latent variable (i.e., ability level) underlies examinee performance, whereas LCA assumes a qualitative latent variable (i.e., class membership) underlies examinee performance. Therefore, we can also use the MRM to investigate our earlier example in which we proposed using LCA to examine the possibility of either two or three distinct performance categories when setting a standard for a test. If a 3-class MRM is championed, three categories of examinee performance are implied to exist, with theta values capturing within-class differences in ability.

As in LCA, the equation for a general C class mixture model (Equation 1) is also used in the MRM. In the MRM, however, the conditional or class-specific probability of a correct response to any item can be expressed as shown in Equation 3.

\[ P(x_i = 1 | c, \theta) = \pi_{ic} = \frac{\exp(\theta - \tau_{ic})}{1 + \exp(\theta - \tau_{ic})} \] (3)

In Equation 3, the conditional probabilities of a correct response are represented by an IRT model (shown here as the Rasch model), where theta (\( \theta \)) is the ability of the examinee and \( \tau_{ic} \) are class-specific item difficulties. The MRM allows within-class correlations between items, unlike LCA, and those correlations are captured by theta (Jiao et al., 2011; Muthén & Asparouhov, 2006; Templin & Jiao, 2012). The MRM also allows examinee ability to differ amongst members of a class, and those within-class differences are also captured by theta (Jiao et al., 2011; Muthén & Asparouhov, 2006; Templin & Jiao, 2012). LCA does not allow for within-class variability; all examinees in a given class are assumed to possess identical ability. Thus, LCA tells us which class an examinee is most likely to be in, whereas the MRM tells us which class an examinee is most likely to be in and an examinee’s relative standing (i.e., theta) within that class (Muthén & Asparouhov, 2006; Templin & Jiao, 2012). If the within-class variance of
theta is constrained to equal zero in the MRM, we will get the same results using LCA as we do with the MRM. Hence, the LCA model is nested within the MRM.

To clarify the similarities and differences between LCA and the MRM, consider the profile plot shown in Figure 4. The graphic illustrates results of the MRM technique based on a five-class solution from data simulated by Jiao et al. (2011). Similar to the profile plot in Figure 3 based on Brown’s LCA results, Figure 4 shows \( \pi_{ic} \), or the probability of a correct response to each item in each class. In LCA, the profile for each class summarizes the probability of a correct response for all examinees within that class. In contrast, the solid lines in Figure 4 only display the probability of a correct response to each item for the average theta level in each class. Variability in item responses within a class in the MRM are captured by the theta values of examinees. The dashed line within each class (shown only for Class 1 in Figure 4) represents the probability of a correct response to each item at one standard deviation above and below the average theta level in that class. As can be seen in Figure 4, the MRM permits variability around the average theta value within any given class.

Jiao et al. (2011) offered an example of how the MRM may be applied as a technique for standard setting. Jiao et al. (2011) simulated data based on results from a bookmark standard setting for the reading portion of a language proficiency test. For the bookmark standard setting procedure, five performance categories were specified for the reading subscale of the test and cut scores were set on the theta scale. In their study, Jiao et al. (2011) simulated theta values for five different classes of simulees on the test to align with the five performance categories from the bookmark procedure. Theta values in the simulated data were assumed to be normally distributed within each class and across
all classes (i.e., the entire simulated sample). The class-specific theta means and standard deviations selected by Jiao et al. (2011) ensured: (a) classes were ordered along the theta continuum (allowing ease of interpretation of theta values across and within classes), and (b) an overlap between adjacent classes that occurred within the same region of the theta continuum as the cut scores derived from the bookmark standard setting procedure. Item difficulties were generated from a standard normal distribution for the middle group (i.e., third performance category). Adjustments were made to these item difficulties to ascertain the class-specific item difficulties for the remaining classes. For instance, adjustments were made so that classes of lower ability had (generally) higher item difficulties and classes with higher ability had (generally) lower item difficulties. These specifications also ensured classes were ordered along the same theta continuum; again, enabling interpretation of theta across all classes.

The generated theta and item difficulties were used by Jiao et al. (2011) to obtain scored responses for simulees, which were then entered as input into the MRM analysis. Specifically, item responses were generated for 10,000 simulees and MRM models with 1 to 7 classes were evaluated based on their fit to the simulated data. Jiao et al. (2011) championed the 5-class model and concluded the simulated results (i.e., the proportion of simulees, the average theta value, and the standard deviation of the theta values) were similar to the values from the bookmark standard setting procedure that were used to generate the data. Jiao et al. (2011) proceeded by obtaining cut scores based on the results of the 5-class model. Jiao et al. (2011) recommended using the theta values associated with the intersection point of adjacent test score distributions as the performance standard if classes are ordered along a continuum, a strategy originally suggested by Hambleton
and Eignor (1980). In this case, the five classes in the championed MRM solution were ordered (because the data were simulated to produce ordered classes), so performance standards were set using the values where theta distributions from adjacent classes intersected. Jiao et al. (2011) compared the findings from their MRM classifications using the simulated performance standards to the classifications of examinees using the bookmark method. The overall classification accuracy of examinees was 86.3%, and ranged from 73.0% to 92.4% for the five groups. Jiao et al. (2011) concluded the classification accuracy of the MRM was relatively high.

**Summary of Data-driven Approaches to Standard Setting**

Now that the various data-driven approaches to standard setting have been described, let’s take a step back and summarize what these classification techniques are generally designed to accomplish. The data-driven approaches to standard setting are used to explore various solutions, each with a different number of groups. All of the data-driven techniques result in classification of examinees into groups based on a pattern of responses to the items of a test. Some authors use the classification results from a data-driven technique as the final determinant of how to group examinees (e.g., Brown, 2007), whereas others choose to use the results to set performance standards on the total score scale using either raw scores (e.g., Sireci, 1995; Sireci et al., 1999) or theta values (e.g., Jiao et al., 2011). As a reminder, studies thus far have only used numerical and statistical classification methods to provide demonstrations of how data-driven techniques could be used to create performance standards. Each method described was merely an example of what standard setting would look like using that approach. A review of the literature
indicated data-driven approaches to standard setting have not been used in practice, so the ramifications of using data-driven techniques in practice are unclear.

The general characteristics of data-driven approaches to standard setting lead to three important considerations when using these procedures. First is the need for large samples of data if a data-driven approach is used to establish performance standards. Additionally, the full range of ability levels needs to be captured on the test in order to make valid interpretations of the results from a standard setting (Sireci, 1995). It is also vital to perform replications to show the stability and validity of the championed solution. Evaluation of repeated samples over time is certainly necessary to ensure classifications are stable and meaningful (Sireci et al., 1999), but replication is also important with data-driven classification techniques because of the general uncertainty behind which is the “correct” solution. Lastly, validity evidence should be collected when using data-driven classification techniques because groups will emerge even if none truly exist. For standard setting purposes, there is a need to make sure classifications are accurate and meaningful; more broadly, validity evidence is stressed with data-driven classification techniques because groups are unobserved. Providing evidence that differences between groups occur in expected ways lends further credibility to the championed solution.

**Advantages and Disadvantages of Data-driven Approaches**

The data-driven approaches to standard setting offer unique attributes to the classification of examinee performance. For instance, the only necessary ingredients to set performance standards using data-driven approaches are participants and data (Sireci, 2001). Data-driven approaches require fewer resources and less time than traditional, panelist-based methods because SMEs and training on a standard setting method are not
needed. Sireci et al. (1999) mentioned the use of cluster analysis as a tool for determining a cut score interval rather than specific point-estimate. A range of values for a potential cut score may offer flexibility to the multitude of interests and factors (e.g., political, social, legal) stakeholders must consider (Sireci et al., 1999).

Another advantage of data-driven approaches to standard setting is the ability to quickly acquire a greater amount of information than is obtained using solely traditional approaches. A particular piece of information obtained exclusively from mixture models (e.g., LCA or the MRM) is the posterior probabilities of class membership for every examinee. In a traditional standard setting, examinees are classified into one group, and one group only. In mixture models, each examinee has a probability associated with his/her membership in each class. The posterior probability of membership in class \( c \), given the responses of an examinee to a set of items \( \mathbf{X} \), is calculated using Equation 4.

\[
P(c|\mathbf{X}) \propto \rho_c \ P(\mathbf{X}|c)
\]

In Equation 4, \( P(\mathbf{X}|c) = \prod_{i=1}^I P(x_i|c) \) specifies the conditional probability of a vector of item responses by an examinee, given class \( c \), as equal to the product of the class-specific or conditional probabilities for each item \( x_i \) in each class \( P(x_i|c) \). The posterior probability of class membership can be estimated for every examinee in every class of an LCA or MRM model using Equation 4. If the posterior probability of membership within a given class is very likely or unlikely (i.e., near 1 or 0, respectively), there is great certainty in the classification of an examinee. Conversely, as the posterior probability of membership within a given class strays from 1 or 0, there is greater skepticism about the resulting classification of that examinee. For example, one examinee might have very distinct probabilities of membership in each class of a 3-class solution.
(e.g., Class1\textsubscript{prob} = 1.00, Class2\textsubscript{prob} = 0.00, Class3\textsubscript{prob} = 0.00), whereas another examinee’s most likely group is less clear (e.g., Class1\textsubscript{prob} = 0.33, Class2\textsubscript{prob} = 0.33, Class3\textsubscript{prob} = 0.33). Clearly, there is greater confidence assigning the first examinee to a class than the second examinee. Other summary information can also be generated in the statistical output of data-driven approaches, such as the proportion of examinees within each class, the proportion of correct responses to each item within each class, and the quality of examinee classification (i.e., entropy). Although some of the summary statistics can also be obtained for traditional approaches to standard setting, the information is not readily available and must be calculated separately. Hence, the researcher may be less inclined to use such information.

Perhaps the most important perceived advantage of data-driven approaches is the potential to bring an “objective” perspective to standard setting, which can be used to combat the frequent criticisms that performance standards created using only SMEs are established arbitrarily and prone to subjectivity (Jiao et al., 2011). However, subjectivity is still involved in the interpretation of findings from data-driven approaches because researchers must 1) identify what they deem to be the most meaningful groups, 2) select the number and nature of groups most appropriate for their standard setting purposes, and 3) determine if and how to establish cut scores using the selected data-driven classification technique (Sireci et al., 1999). It is also important to keep in mind that data-driven classification techniques will always partition examinees into groups, regardless of whether they actually exist. The inevitable creation of groups may lead to situations in which artificial differences are found that are not meaningful. Consider a situation where a data-driven technique provides a solution with a “construct-irrelevant” class. For
instance, a class with item difficulties close to obtaining a correct response by chance alone in LCA or the MRM might represent an unmotivated class, not necessarily one low on the construct in question. Again, the uncertainty of whether groups characterize valid and meaningful differences among examinees stresses the importance of replication and validation.

Another issue with data-driven approaches as a primary tool for standard setting is the potential for proficiency groups to emerge that are unordered. The ultimate goal of standard setting is to place examinees into groups that order scores along a continuum, separating *Developing* from *Proficient* from *Advanced* performance, for example. Each study described in the literature review that employed a data-driven approach to standard setting found ordered performance groups—a fortunate finding, but surely not always the case in practice. There will likely be occasions where data suggest performance groups are not ordered for a test, as shown in the profile plot displayed in Figure 5. The response profiles are not parallel, suggesting students differ *qualitatively* rather than *quantitatively* and should not be ordered along a continuum (i.e., a single continuous score). That is, the groups appear to be nominal and fail to describe quantitative differences between examinees. Standard setting may not be an appropriate course of action if we find non-parallel response profiles in our data because it implies groups exist, but the groups are nominal, not ordinal. There is also the potential for mismatch between the number of groups expected to emerge from the population and the number of groups found using a data-driven classification technique (Templin & Jiao, 2012). For instance, policymakers may desire three performance groups, but LCA indicates a 2-class solution fits the data about as well as a 3-class solution. In this situation, the researcher has a difficult decision
about whether to consider the classification from SMEs or the LCA model as the most appropriate interpretation. Such a decision could easily be construed as a “subjective” choice on the part of the researcher.

Perhaps it is the advantages and disadvantages outlined above that have prompted some authors to suggest results from data-driven classification techniques should not be used to set performance standards on a test. Instead, the results could be used to inform SMEs while they are conducting traditional standard setting studies, to inform stakeholders who are creating performance categories, or to provide validity evidence for other standard setting techniques. Sireci et al. (1999) advocated the use of classification techniques as a way to augment traditional standard setting studies. They did mention data-driven classification techniques can be used to set standards, however, if a great deal of external validity evidence is available. If data-driven approaches are used as the primary source to set performance standards, SMEs should still be heavily involved in the interpretation of results (Sireci et al., 1999).

**Integration of Traditional and Data-driven Approaches to Standard Setting?**

It appears there is no easy solution to conducting a standard setting. It can be difficult just to gather a group of SMEs to serve on a standard setting panel, let alone one representative of all stakeholder interests (Bunch, 2012). Even if a diverse and qualified set of judges is assembled, a bounty of time and resources must be expended to develop appropriate performance category descriptors, train SMEs, and have them assign ratings regarding item difficulty or examinee ability (Bunch, 2012; Hambleton & Pitoniak, 2006). Even still, the results of a traditional standard setting remain open to criticism that they are too subjective or arbitrary to merit value. Although data-driven approaches to
standard setting are intended to address some of the issues inherent with traditional approaches, problems are prevalent with their use as well. Numerical and statistical procedures require data to be available for analysis in order to establish performance standards, which is not always an ideal situation for testing (in higher education or other fields). Additionally, all data-driven approaches to standard setting will classify test takers into groups, but the techniques remain oblivious to whether those groups are actually meaningful. It is the researcher’s discretion to make a decision about which solution fits the data best and also forms a practically reasonable grouping of examinees.

The results from data-driven approaches may also conflict with the number and nature of performance categories desired by policymakers and other stakeholders. So even though data-driven approaches to standard setting seem to promote objectivity, they may not be immune to the same subjectivity required for traditional approaches.

Both traditional and data-driven approaches have the capacity and intent to produce quality results, but each possesses difficult-to-overlook limitations. The field has long ago come to a common understanding there is no “true” cut score that exists for a test, awaiting human discovery (Cizek, 1993, 1996a, 1996b, 2012; Hambleton, Pitoniak, & Coppella, 2012; Jaeger, 1989; Kane, 1994a; Livingston & Zieky, 1982; Sireci et al., 1999; Zieky, 2012). Researchers in the field are also beginning to realize there is no correct or “right” standard setting method (Cizek, 2012; Zieky, 2012). Perhaps neither traditional nor data-driven methods should be used as the sole means to set performance standards on a test. There has always been an element of judgment inherent to setting standards—a qualitative sort of analysis—that can be especially useful and complementary of the quantitative component of the process. Likewise, the results of
data-driven approaches to standard setting are incapable of proper utility in isolation of human interpretation; that is, without context and the role of judgment (particularly from SMEs) regarding test content and the examinee population, performance standards created using only advanced numerical and statistical algorithms seem insufficient. Instead, maybe the best course of action is to integrate the work done by policymakers, stakeholders, and SMEs—each of which provide human judgment/context to the meaning of scores and groups—with the statistical richness of data-driven classification techniques.

**Integrated Approaches to Standard Setting**

Integrated approaches to standard setting combine judgment from SMEs with statistical analyses to classify examinees in a way that is based on integrated information from both traditional and data-driven approaches. Although there have not been any applications of integrated approaches to set performance standards to date, the details of two demonstrations of integrated techniques—each of which utilized mixture models—are described in the next two sections.

**Augmented LCA.** There are two main differences between general LCA and the augmented LCA proposed by Templin et al. (2007). First, information external to the data is used to inform model estimation in the augmented LCA. In the general LCA, examinees are assigned into classes based solely on their pattern of responses to items of a test. In the augmented LCA, SMEs’ classifications of examinees into groups (obtained from a traditional, examinee-centered standard setting) accompanies the examinees’ item response data as input into the analysis; that is, the data used for an augmented LCA come from a pair of sources. Examinees are classified into groups on the basis of
evaluator’s ratings regarding a student’s ability *and* on the item response pattern of each student. Second, there is not an exploration or comparison of the potential number of groups that best models the data in an augmented LCA. Rather, the number of groups is decided upon by stakeholders before the augmented LCA procedure is conducted, and only a solution with that number of classes is fit to the data. If only two performance groups are decided upon by stakeholders, SMEs’ evaluations of examinee ability are only available for those two groups. There is no way to compare solutions in the augmented LCA using anything other than two groups because the information collected from SMEs involved only two groups. The results from an augmented LCA produce summary statistics similar to those generated by the general LCA. Both LCA methods provide the researcher with estimates of the proportion of examinees within each class (i.e., class weights), class-specific item difficulties, and the probability of each examinee’s likelihood of being in each performance group.

An applied example of the augmented LCA was demonstrated by Templin et al. (2007) for end-of-grade tests to assess reading and mathematics in Kansas schools. Five proficiency groupings were already established by the state of Kansas for the tests, but the creation of new versions of the test prompted a need for updated performance standards. Templin et al. (2007) described the augmented LCA as a three-part process: (1) Collect rating information from SMEs about examinees, (2) Conduct LCA and obtain item- and class-level parameter estimates, and (3) Use the parameter estimates to assign examinees to performance levels. Templin et al. (2007) collected ratings from teachers familiar with students’ academic work using two examinee-centered approaches to standard setting. The teachers in Kansas were allowed to use either the contrasting groups
or the borderline group approaches by assigning each student to either one or two performance categories. For instance, if there were five performance categories, teachers could choose to classify a student in either one of the five categories (the contrasting groups approach) or into two adjacent categories (the borderline groups approach). Out of 18,519 students who took the test, 2,626 received ratings from teachers or administrators. Templin et al. (2007) used $r_{sc}$ as an indicator of the teacher’s performance category decision $r$ for student $s$ in category $c$. In the contrasting groups method ($n = 1,953$), students were only assigned to one group by teachers using the rule: $r_{sc} = 1$ if rating is $c$, otherwise 0. An example of a set of ratings for a student deemed to be in the third of five performance categories is $r_s = [0, 0, 1, 0, 0]$. For the borderline group method ($n = 673$), students were classified into two adjacent groups by teachers which was represented as: $r_{sc} = 0.5$ if rating on the border of category is $c$, otherwise 0. For example, a student on the border of the second and third performance categories is given two 0.5 ratings of class membership, or $r_s = [0, 0.5, 0.5, 0, 0]$.

In order to classify students into performance groups using the augmented LCA, the ratings from teachers must be integrated into the equation we used earlier for the general LCA. Much of the equation for the augmented LCA used by Templin et al. (2007) is very similar to that used for the general LCA, with one notable addition: the ratings provided by SMEs ($r_{sc}$) must also be included. The augmented LCA is defined by Equation 5,

$$P(x_i = 1) = \sum_{c=1}^{C} r_{sc} \left[ \rho_c * P(x_i = 1|c) \right]$$

(5)

where $P(x_i = 1|c) = \pi_{ic} = \frac{\exp(-\tau_{ic})}{1+\exp(-\tau_{ic})}$, $\pi_{ic}$ is the probability of a correct response to item $i$ in class $c$, just like in the general LCA (Equation 2), and $\tau_{ic}$ represents the item
difficulty on the logit scale for item \( i \) in class \( c \) (i.e., within-class item difficulties, which was also represented in Equation 2).

The ratings from the teachers are also used in the calculation of the posterior probabilities of class membership for each examinee. Recall, the probability of being assigned to each of the performance categories is assigned to students after LCA is estimated. In the Templin et al. (2007) example, five posterior probabilities of membership were obtained for each student who took the Kansas reading and mathematics test after the augmented LCA was conducted. The posterior probabilities of class membership in the augmented LCA were modeled in the same way shown earlier for the general LCA (Equation 4), but with the teacher ratings \( (r_{sc}) \) included, as shown in Equation 6.

\[
P(c|Χ) \propto r_{sc}(\rho_c * P(Χ|c))
\]

Classification of students into performance groups using the augmented LCA was based on their highest posterior probability. (Templin et al. [2007] did demonstrate how cut scores on the total score scale could be derived through a Monte Carlo simulation study, but they did not report the cut scores.) Classifications made using the augmented LCA were compared to classifications based on other methods. Specifically, ratings from the contrasting groups and borderline group methods were used to create cut scores and classify examinees—as is typically done with traditional approaches to standard setting. Templin et al. (2007) concluded the augmented LCA compared favorably to the two traditional standard setting methods, which provided validity evidence for the use of an integrated approach to standard setting.
**Adapted MRM.** The adapted MRM was designed by Templin et al. (2008) as another integrated technique to incorporate traditional and data-driven approaches to standard setting. There are a few similarities between the general MRM and the adapted MRM, namely the statistical output produced from each technique. Specifically, class weights, theta (or IRT estimates of ability) for each examinee, and class-specific item difficulties are obtained for both models. The posterior probability of class membership for each examinee in each class is also obtained, just as in LCA, augmented LCA, and the general MRM.

The adapted MRM differs from the general MRM in some regards as well, though only a conceptual description is provided here because neither model is used in the present study. The key difference between the two MRM models is the use of the borderline or contrasting groups classifications from SMEs as priors for class membership for each examinee in the adapted MRM. Additionally, there is no exploratory process to find the correct number of classes with the adapted MRM, which is also the case for the augmented LCA. The number of classes specified by the user of the adapted MRM corresponds to the number of performance categories mandated by stakeholders. The adapted MRM also estimates a correlation between class membership and theta. Templin et al. (2008) asserted the class a student is assigned to should be “significantly correlated with the ability of the student” (p. 389). In other words, examinees with higher ability are expected to be assigned to more advanced performance categories than those with lower ability. Another feature that differentiates the adapted MRM from the general MRM is the ability to calculate the Kullback-Leibler Divergence (KLD) index. The KLD index conveys each item’s ability to discriminate between
students categorized into different performance groups and helps test developers select items that best discriminate between classes (Templin et al., 2008). Templin et al. (2008) fit the adapted MRM to the same data used in their illustration of the augmented LCA. They obtained a correlation of 0.73 between ability level and performance category, and also demonstrated how the KLD index could be used to select items to create a shortened version of the test that maximally discriminates between different performance categories.

**Summary of Integrated Approaches to Standard Setting**

The integrated approaches to standard setting combine both expert judgment and statistical analysis to classify examinees into performance groups. Two integrated approaches have been proposed. The augmented LCA (Templin et al., 2007) incorporates teachers’ ratings of student proficiency levels as prior information to inform a general LCA, which assigns examinees into groups based on item response profiles. Results showed classifications made using the augmented LCA compared favorably to two examinee-centered methods of standard setting. Later, Templin et al. (2008) adapted their integrated approach by using teacher judgment regarding student performance as prior input into a MRM. Results from the adapted MRM provide information about differences in ability between groups as well as the relative standing of students within the same group. Additionally, the strength of items to discriminate between performance categories can be calculated via the KLD index, in conjunction with the adapted MRM.

The clear advantage of integrated approaches is the means to create performance groups using multiple sources of information (Templin & Jiao, 2012). Researchers may make more informed decisions about student performance by pairing judgment from
SMEs with the results of statistical analyses. The integration of traditional and data-driven methods also addresses common criticisms of standard setting approaches. Traditional methods are often considered too subjective because they rely solely on human judgment. Results from data-driven methods may produce groupings that deviate in number and in nature from the performance categories created by policymakers. Ultimately, the findings from a traditional or data-driven classification technique have shortcomings. Using information compiled from both methods demonstrates an effort to integrate contrasting, unique perspectives and avoids reliance on one type of approach when making interpretations.

There are some concerns associated with integrated approaches to standard setting as well. For one, the number of performance groups must be established by policymakers and stakeholders before analytical techniques are applied to the data. As such, the fit of only one model can be evaluated when using an integrated approach. In contrast, purely data-driven approaches can be used to explore a number of different models. The lone model tested in an integrated approach is constrained such that the number of groups equals the number of performance categories established prior to data analysis. If the number of established performance categories does not align well with the data, the fit to the data of the lone model tested as part of the integrated approach is likely to be poor. In this case, stakeholders are faced with a difficult decision about how to proceed, given conflicting evidence from each component of the integrated approach (i.e., results from the traditional versus data-driven portions).

Second, to estimate the model parameters, complex computational packages must be created by the user or adapted from another researcher. The researchers who
formulated the augmented LCA devised their own programming code to perform the analyses using statistical software. Further, Templin et al. (2008) used a fully Bayesian approach with the adapted MRM. Working from a fully Bayesian approach may be impractical due to a lack of expertise for many practitioners. Thus, a lot of technical prowess is needed to appropriately apply the integrated approaches to standard setting. The integrated approaches, as currently constructed, may be too complicated for everyday use.

So far, integrated approaches to standard setting have only been used with examinee-centered methods. As outlined earlier in this chapter, there are also concerns associated with the use of examinee-centered methods for standard setting. First, judges must have enough familiarity with the testing population to expertly rate expected levels of performance for each examinee (Hambleton & Pitoniak, 2006). More importantly, the validity of interpretations made from the results of examinee-centered methods may be threatened if judgment by SMEs includes factors unrelated to student ability (Cizek, 1996; Jaeger, 1989). Maybe a teacher who is participating in a standard setting knows certain students give really good effort on all assignments. Although those students may not always get the correct answer, they do always try hard and the teacher feels compelled to reward the effort of those students. Unfortunately, work ethic is not necessarily related to examinee ability. In the context of standard setting, judgments from SMEs about matters unrelated to examinee ability represent construct-irrelevant variance—a major threat to validity (Benson, 1998; Cronbach & Meehl, 1995; Kane, 2001b; Messick, 1995). An important component of examinee-centered methods is the
judgment SMEs provide about examinees. A reliance on human judgment, however, may also introduce error into (or diminish the validity of) the standards being set.

Ultimately, there is a lot of promise for the use of integrated approaches to standard setting; however, there are legitimate concerns with their current structure. If integrated approaches are to become a viable proposition for future standard settings, there likely need to be modifications considered that make them more accessible for everyday use and improve confidence in the validity of the standards and groupings that result.

**Purpose of the Current Study**

The purpose of the current study is to examine the results from various standard setting approaches to assess how each fare relative to the other methods. Performance standards were set on the Ethical Reasoning Identification Test-XA (ERIT-XA) using four different classification methods. The first standard setting method was a traditional approach: the modified Angoff. Three performance groups were specified by stakeholders for the ERIT-XA prior to the traditional standard setting: Developing, Proficient, and Advanced. Justification for the selection of the modified Angoff, details of the performance category descriptions, and an overview of how the standard setting was implemented will be provided in Chapter Three.

Two data-driven approaches to standard setting were used to classify examinees into performance groups as well. The sample for each data-driven method of classification included students who took the ERIT-XA. The details of data collection and testing context will be described in Chapter Three. The general LCA described previously was used as one of the two data-driven approaches to set performance
standards on the ERIT. Recall, parameters of the general LCA model are free to vary; there are no constraints placed upon the difficulty of items within a given class, for instance. The second data-driven model used for classification was the ordinal LCA. The ordinal LCA differs from the general LCA in that constraints are placed on the parameters of the model. Constraints were put on the item difficulty values for each class during estimation of the LCA model, such that more advanced performance groups had higher item difficulties (i.e., items will be easier). Essentially, the ordinal LCA model constraint can be described as: \( \pi_{i3} > \pi_{i2}; \pi_{i2} > \pi_{i1} \), where \( \pi_{ic} \) represents the within-class item difficulty for item \( i \). Constraints were made such that only items that were unordered in the general LCA were forced to be ordered in the ordinal LCA, which was the recommended methodology proposed by Croon (1990; 2002). The constraints made during estimation in Mplus were in accordance with the manner described by Finch and Bronk (2011), which requires all classes to be ordered.

Lastly, an integrated approach to standard setting was performed using the same sample of students. The proposed integrated approach incorporated SMEs’ evaluations of test content with a data-driven classification technique. Specifically, item ratings assigned by SMEs during the modified Angoff were used to inform a LCA as part of an integrated approach to standard setting, referred to henceforth as the Angoff LCA. The Angoff LCA had even more constraints than the ordinal LCA. Ratings provided by SMEs during the modified Angoff standard setting were used to constrain the item difficulties for each item within each class. All classes were still ordered, just as in the ordinal LCA, but a value (adopted from the modified Angoff) was used to separate performance groups (i.e., as a cut score). Specifically, the item-difficulties were ordered to align with the ratings
from SMEs for each item. An example of the modifications made for the Angoff LCA model constraint is: $\pi_{i3} > MA32_i > \pi_{i2} > MA21_i > \pi_{i1}$, where $\pi_{ic}$ represents the within-class item difficulty for item $i$, $MA32_i$ is the probability (median across all SMEs) of a correct response for examinees at the border of classes 3 (Advanced) and 2 (Proficient) for item $i$, and $MA21_i$ is the probability (median across all SMEs) of a correct response for examinees at the border of classes 2 (Proficient) and 1 (Developing) for the same item.

So far, integrated approaches to standard setting have only been used with examinee-centered methods. In the current study, results from a test-centered method were used to inform a LCA model as part of an integrated approach to standard setting. There are no other examples of an integrated approach to standard setting that combines the results of a test-centered method with a data-driven method of classification. The judgments obtained from SMEs participating in a test-centered method may be more trustworthy than those from an examinee-centered method because test content is emphasized as part of the rating process (Cizek, 2012; Jaeger, 1989). It may also be more practical to gather a group of SMEs who are adept at evaluating test content than it is to find a group of experts who are intimately familiar with an actual group of test takers and test content (Hambleton & Pitoniak, 2006).

**Research Questions**

Of primary interest is whether the classification of students differs based on the standard setting methods used. Such an analysis will shed light on whether the various approaches to standard setting lead to differential results (e.g., percentage of examinees
classified into each group). The following research questions (RQs) were used to guide the current study:

1. Which model is championed as the best fit to the data for the general LCA and the ordinal LCA? Do the general and ordinal LCA each indicate a consistent number of classes as the best solution? For example, is a 3-class solution suggested as the best fit to the data for each LCA model (i.e., general LCA, ordinal LCA)?

2. How do the results from the championed general LCA model, championed ordinal LCA model, and Angoff LCA compare? Are the class-specific item difficulties similar or different?

3. Across the three approaches involving statistical techniques, which model is championed as the best fit to the data?

4. Are similar percentages of students classified into each performance group using the various approaches to standard setting (i.e., modified Angoff, general LCA, ordinal LCA, and Angoff LCA)?

5. Do the relations between classifications and external variables depend on the standard setting method used? Also, do the relations with external variables align with expectations, thus providing validity evidence for the standard setting method used?
CHAPTER THREE

Methods

The methods chapter will begin with background and a description of the measure used in the current study. An overview of the traditional standard setting method implemented in the current study will be presented next. The overview will include details about how the selected method was chosen, qualifications of SMEs who participated in the traditional standard setting, the performance categories created for the traditional standard setting, and the training process of SMEs for the selected method. A description of the data collection process, participants, and statistical analyses will follow. Focus will then shift to the data-driven and integrated approaches to standard setting used in the current study. Specifics about the similarities and differences between the constraints for the data-driven and integrated models will be provided. Next, a discussion of the estimation method used will be presented as well as details regarding model-data fit comparisons. The set of five RQs of interest in the current study will follow, with a plan for how each set of RQs will be analyzed. Lastly, validity analyses and variables used as external sources of validity evidence for the current study will be described.

Measures

Ethical Reasoning Identification Test-XA (ERIT-XA). The ERIT-XA was developed in 2012 by ethical reasoning experts and faculty at James Madison University (JMU) as part of the institution’s Quality Enhancement Plan, “The Madison Collaborative: Ethical Reasoning in Action” (MC). The MC was designed to enhance ethical reasoning as a priority for undergraduate students at JMU by connecting related
activities and curriculum (The Madison Collaborative, 2013). The MC has developed Eight Key Questions (8KQ) that are introduced to students during freshman orientation. A vast majority of JMU students receive educational programming in ethical reasoning at the beginning of their first year via the It’s Complicated intervention (Smith, Fulcher, & Pyburn, 2015). During It’s Complicated, all first-year students are split into groups to discuss an ethical reasoning scenario using the 8KQ. The activity is designed to emphasize the importance of ethical reasoning to students as well as familiarize them with the 8KQ (Smith et al., 2015). The 8KQ (shown in Table 2) consist of single-word terms that suggest ethical considerations for personal or group reflection before making a decision (Smith et al., 2015). There are also other opportunities for students at JMU to experience ethical reasoning programming outside of It’s Complicated. For example, some faculty have participated in professional development workshops facilitated by the MC and have begun to integrate the 8KQ framework into their courses (Smith et al., 2015). Students may also receive additional instruction on the 8KQ and other ethical reasoning interventions throughout their careers at JMU, contingent upon their major and general education courses (Smith et al., 2015; Smith, Pyburn, & Ames, 2016; The Madison Collaborative, 2013).

The ERIT-XA is one of a battery of assessments used by the MC to assess students’ ethical reasoning skills and attitudes at JMU (Smith et al., 2015). The ERIT-XA is a 50-item test comprised of 42 multiple-choice questions and two additional, 4-item testlets. Thus, there are 42 dichotomously scored items on the test and 8 additional items that could either be scored polytomously, resulting in two items each scored on a 0-4 scale, or dichotomously, resulting in eight items each scored on a 0/1 scale. Each item on
the ERIT-XA consists of a scenario where an ethical dilemma is presented to the test taker. For example, item 2 of the ERIT-XA reads, “At a soup kitchen operating on tax dollars, Pete serves a meal to the homeless every week because he thinks every human being is entitled to at least one good meal a day.” Students are provided a brief description of the 8KQ (shown in Table 2) at the top of each page when they take the ERIT-XA. Examinees are instructed to select which of the 8KQ is most consistent with the ethical decision or rationale, which, in item 2, was the decision made by Pete about feeding the homeless. Of the five cognitive student-learning outcomes (SLOs) created by the MC, the ERIT-XA is aligned to two: (1) When given a specific decision and rationale on an ethical issue or dilemma, students will correctly identify the Key Question most consistent with the decision and rationale; (2) Given a specific scenario, students will identify appropriate considerations for each of the eight Key Questions (Smith et al. 2015).

Factor models and item response theory (IRT) models have been fit to the responses to the ERIT-XA to investigate its dimensionality and whether the items on the testlets relate to one another beyond that expected by the latent trait(s) (Smith et al., 2015; Smith et al., 2016). The dimensionality of a test is pertinent to standard setting in order to determine whether a total score or multiple subscale scores are more suitable for classification purposes (Sireci, 1995; Sireci et al., 1999). The fit of the data to a one-factor (or unidimensional) model and an eight-factor model (one factor for each of the 8KQ) has been evaluated for various administrations of the ERIT-XA items over the past few years. The one-factor (or unidimensional) model has repeatedly shown superior fit to the data compared to the eight-factor model (Smith et al., 2015; Smith et al., 2016). Local
misfit of the one-factor and eight-factor models has also been examined by Smith et al. (2015) and Smith et al. (2016). Correlation residuals did not yield patterns in support of an eight-factor structure (i.e., the majority of the local misfit was not due to the Key Questions). There was also no excessive dependency among the items associated with the same testlet to treat scoring of those items as polytomous instead of dichotomous. Thus, the one-factor model was championed as the best fit to the data (Smith et al., 2015; Smith et al., 2016). Because evidence indicates the ERIT-XA represents one latent construct, setting a standard on the total score is suitable.

**Traditional Standard Setting**

In the present study, performance standards are set on the ERIT-XA using a variety of approaches and the results obtained from each method are compared. One of the methods used to set performance standards on the ERIT-XA was a traditional standard setting procedure. Hambleton and Pitoniak (2006) identified nine steps as typical of any traditional standard setting procedure. For space constraints, only the most pertinent of the nine steps delineated by Hambleton and Pitoniak (2006) are described in this section—with particular emphasis on how they were accomplished in the current study. All nine steps of the traditional standard setting procedure are displayed in Figure 6 as a point of reference for when each occurred in the context of the current study.

**Modified Angoff.** Aligning with Figure 6, three steps were undertaken prior to conducting the traditional standard setting, the first of which was selecting an appropriate method. Central to the selection of a standard setting method are the types of items on the test and the amount of time/resources available (Hambleton & Pitoniak, 2006). Consideration of the types of items on the test is particularly important because different
standard setting methods are appropriate for different kinds of items. The modified Angoff was chosen for the traditional standard setting on the ERIT-XA because it is often used with multiple-choice exams (Hambleton & Pitoniak, 2006; Kane, 1994b; Mehrens, 1994), has produced reliable ratings from SMEs (Cizek, 1996b), and has a good balance of technical and practical application (Mills & Melican, 1988).

Two important considerations typically surface regarding the composition of a standard setting panel: 1.) How many raters should be chosen, and 2.) What qualifications are sufficient of an “appropriate” rater? An extensive literature review by Brandon (2004) showed 10 to 20 raters is an appropriate and effective number for the modified Angoff. Raymond and Reid (2001) suggested between 10 to 15 raters to ensure reliable judgments are made. Hambleton and Pitoniak (2006) recommended 15 to 30 raters as an acceptable number in order to enable representation of all stakeholders and maximize stability of results. A group of 10 raters participated in the ERIT-XA standard setting. The raters represented various branches of JMU—including faculty, administration, staff, and developers of the ERIT-XA—and were offered a stipend by the MC for their efforts.

One of the most important aspects of a standard setting is creating clear descriptions of the performance categories in which examinees are to be classified. Descriptions of performance categories are usually established in advance of a standard setting and should embody the student-learning objectives of a program (Ricker, 2006). Prior to the standard setting, an ethical reasoning SME and facilitators of the ERIT-XA standard setting convened in the Fall 2016 semester to discuss the number and nature of the performance categories that should be used for the ERIT-XA. Three categories of student performance were identified for classification purposes: Developing, Proficient,
and Advanced. Complete descriptions of the evaluative criteria for each performance category are displayed in Table 3.

The modified Angoff standard setting for the ERIT-XA was conducted in February of the Spring 2017 semester. An agenda listing all activities of the ERIT-XA standard setting is available as a supplemental guide in Figure 7. At the beginning of the standard setting workshop, SMEs were oriented to the two MC cognitive SLOs assessed by the ERIT-XA. Descriptions of performance categories were also presented to SMEs. It is important SMEs share a common understanding of the evaluative criteria used for each performance group in order to achieve consistency in ratings. An operational definition of what constitutes “minimally proficient” and “minimally advanced” examinee performance was discussed with SMEs because these two groups represent the borderline students that were rated on each item throughout the modified Angoff. The rationale behind standard setting was then explained and SMEs were trained on making ratings using the modified Angoff procedure. A practice session was conducted using three items from a previous version of the ERIT-XA to acclimate SMEs to the rating process.

Following training, SMEs recorded their ratings for “minimally proficient” and “minimally advanced” students. A graphic was displayed to SMEs to provide a visual representation of the groups they were rating (see Figure 8). The aim was to make explicit the two groups of borderline examinees that should be considered by SMEs when making their ratings. Chromebooks or personal laptops were used by SMEs to submit electronic ratings in Qualtrics. Each item and the eight response options (i.e., the 8KQ) appeared in the Qualtrics survey. SMEs each individually progressed through the ERIT-XA item-by-item without discussing the test with their peers. SMEs were instructed to
provide a judgment regarding the number of minimally proficient and minimally advanced examinees (out of a hypothetical 100) they expected would correctly answer that item. SMEs were allowed to make ratings on a scale from 0-100 using a slider created in the Qualtrics survey (Figure 9). The process was repeated for each item on the ERIT-XA. Two rounds of ratings were conducted for the ERIT-XA standard setting.

As part of the modified Angoff procedure, SMEs received feedback between rounds about their individual ratings and the panel’s overall ratings. Each rater received an itemized document with their ratings for both borderline groups as well as the performance standards that individual set. The facilitators of the standard setting aggregated the panel’s ratings and displayed the group’s cut scores in terms of raw items correctly answered and percent-correct on the ERIT-XA. After the data were displayed, SMEs gathered to discuss their ratings from that round in small groups of five, which were led by a facilitator of the standard setting. The ten items with the largest variability in ratings were highlighted and SMEs from each group shared their rationale and considerations when making ratings on those items. “Impact” data were also presented after the second round to show SMEs the percentage of examinees classified into each group. Specifically, the percentage of examinees classified into the Developing, Proficient, and Advanced performance categories were presented to SMEs to demonstrate the implications, or “impact,” of the current performance standards.

The feedback and discussion sessions were ultimately designed to allow SMEs to compare their ratings with other members of the panel. SMEs are not obligated to change their ratings after receiving feedback or engaging in discussion, but often do so as they become aware of errors or misconceptions they made initially (Hambleton & Pitoniak,
To establish cut scores, ratings from the final round were summed across all 50 items of the ERIT-XA for each SME. Either the mean or median of the summed SME ratings is generally used as the performance standard for each round of a modified Angoff procedure (Hambleton & Pitoniak, 2006). The median of the summed SME ratings after the second round was used as the final performance standards separating Developing/Proficient and Proficient/Advanced students. All results pertaining to the modified Angoff are provided in Chapter Four.

**Procedures and Participants**

The data used for analysis were collected from students who took the ERIT-1D, ERIT-1E, or ERIT-XA at a midsized, mid-Atlantic university. The same items were administered on the ERIT-1D, ERIT-1E, and ERIT-XA, but each test was either preceded or followed by a different set of tests. For convenience, the term ERIT-XA—which is the presently-used version—will be used to refer to the test throughout this manuscript. Data were collected from two testing contexts, both of which were supervised by at least one proctor. Date were primarily collected across seven Assessment Day testing sessions, spanning Fall 2013 to Fall 2016 ($n = 3,316$). Assessment Day is an institution-wide, mandatory testing session in which a variety of general education tests are administered to students at the university. Students attend Assessment Day twice during their academic career; once as entering freshmen in their first fall semester and, again, during the spring when they have amassed 45 to 70 credit hours. On each Assessment Day, students are exempt from classes and randomly assigned to rooms to participate in a two- to three-hour testing session. Random assignment of students to particular cognitive and noncognitive assessments is done using the last three digits of their unique student
identification number. The assessments are taken in what is considered a low-stakes context because students are neither penalized nor rewarded for their performance.

Data were also collected from students who took the ERIT-XA at two time points during the Fall 2016 semester as part of research for Madison Collaborative (MC) at JMU. Unlike Assessment Day, the status of students who took the ERIT-XA during MC testing was variable. That is, students who took the ERIT-XA as part of MC testing were not all Freshmen or Sophomores; their current academic year at JMU varied. The students who took the ERIT-XA as part of MC were tested at the beginning of the Fall 2016 semester before participating in ethical reasoning training (pretest), and again toward the end of the semester after receiving training (posttest). For purposes of the current study, only the posttest responses from MC testing were retained for analysis ($n = 242$). The combined Assessment Day and MC data set used for analyses included 3,558 students.

Of the 3,558 students included in the analyses, 131 had missing responses. Listwise deletion of cases with missing data makes a stringent assumption that all data are missing completely at random (MCAR), or that the missing values on a variable are independent of other observed variables (Rubin, 1976). If missing data are not MCAR, listwise deletion may generate biased parameter estimates and inaccurate conclusions (Acock, 2005; Enders & Bandalos, 2001; King, Honaker, Joseph, & Scheve, 2001). Even if data are MCAR, listwise deletion reduces power, or the ability to find an effect if one exists in the population (Enders & Bandalos, 2001; Myers, 2011). Because the reason(s) for missing data were unknown, students with missing responses ($n = 131; 3.6\%$) were retained in the dataset. A majority of the 131 students who had missing responses failed
to respond to only one item \((n = 91; 69.5\%)\). Another 16 students \((12.2\%)\) were missing responses to only two items. Only two students were missing data for more than 10 items—one for 11 items and another for 13—though they still responded to about 75% of items on the test. All missing responses were scored as incorrect answers and included in the total score on the ERIT-XA. The number of students who took the ERIT-XA at each testing administration is displayed in Table 4. In the current sample, demographic analysis revealed 60% of students identified as female and 78% were White\(^1\).

The data set consisting of 3,558 student responses to various administrations of the ERIT-XA was used in many ways in this study. The data were used as input into the data-driven and integrated approaches to standard setting. Additionally, the data were used during the modified Angoff traditional standard setting to provide SMEs an illustration of how students actually performed on the ERIT-XA. For instance, “impact” data provided to the raters during the modified Angoff procedure were based on classifications of students into performance categories using their total scores on the ERIT-XA and the standards that were set by the raters. The data were also used to set performance standards for the data-driven and integrated approaches in the current study.

**Data-driven and Integrated Standard Setting Approaches**

There are three standard setting approaches explored in the current study that use statistical techniques: the general LCA, ordinal LCA—each of which are purely data-driven approaches—and Angoff LCA, which is an integrated approach. The same data set described above containing responses to the ERIT-XA is used with all three models. The data-driven standard setting models are described below, followed by the integrated model and subsequent sections on model estimation and model fit.
**General LCA.** The general LCA was used as one of two data-driven approaches to set performance standards on the ERIT-XA. In the general LCA, parameters of the model are free to vary; for instance, there are no constraints placed upon the difficulty of items within a given class. Item response profiles derived from the general LCA may be parallel (i.e., ordered) or non-parallel (i.e., unordered), depending on the nature of the groups formed as a result of the analysis. For the current study, examination of the performance groups that emerged using the general LCA was accomplished by evaluating the fit of various solutions to the data (e.g., 1-class, 2-class, 3-class, 4-class) until convergence issues were encountered (details regarding convergence issues will be described in the Estimation section). Of all the solutions that satisfactorily converged, one was championed as the best-fitting general LCA model to the data (using the model fit indices described later in the chapter). The championed general LCA model was used for comparison to the best-fitting model from other standard setting techniques employed as part of the current study.

**Ordinal LCA.** The ordinal LCA differs from the general LCA in that constraints are placed on the parameters of the model. Constraints are put on the item difficulty values for each class during estimation of the ordinal LCA model, such that more advanced performance groups will have higher item difficulties (i.e., items will be easier). Specific restrictions on the item difficulty values were not made, just a constraint to force classes to be ordered. Constraints were only made to those items that emerged as unordered in the corresponding general LCA model, as recommended by Croon (1990; 2002). For example, in the 3-class ordinal LCA solution, the within-class item difficulty for unordered items of the 3-class general LCA solution was constrained to be higher for
all examinees in Class 3 (Advanced) than Class 2 (Proficient). Similarly, the item difficulty for unordered items of the 3-class general LCA solution was constrained to be higher for all examinees in Class 2 (Proficient) than Class 1 (Developing). Essentially, the ordinal LCA model constraint is: \( \pi_{i3} > \pi_{i2} > \pi_{i1} \), where \( \pi_{ic} \) represents the within-class item difficulty for item \( i \). The constraint forces classes to be ordered; the third class has a higher item difficulty (i.e., the item is easier) than the second class, which in turn has a higher item difficulty than the first class (Croon, 1990; Croon, 2002; Finch & Bronk, 2011). Just like the general LCA, evaluation of the performance groups that emerged using the ordinal LCA was accomplished by comparing the fit of various models to the data. Solutions of the ordinal LCA model (e.g., 1-class, 2-class, 3-class, 4-class) were fit until convergence issues were encountered. One solution was championed as the best-fitting ordinal LCA model to the data and was compared to the championed model from the other standard setting techniques.

**Angoff LCA.** As previously described, only one model can be considered when using an integrated approach to standard setting. For the Angoff LCA, only a 3-class model was assessed for fit to the data because ratings from SMEs who participated in the modified Angoff were used to classify examinees into three performance groups. The Angoff LCA has even more constraints than the ordinal LCA. The three classes are still ordered—just as in the ordinal LCA—but the specific ratings provided by SMEs from the modified Angoff are used as constraints to separate performance groups on each item. Again, constraints for the Angoff LCA were only made to items in the corresponding general LCA model (i.e., 3-class solution) that were not aligned with the item difficulty values specified by SMEs. Specifically, the Angoff LCA model constraint is: \( \pi_{i3} > \)}
MA32, > πi2 > MA21, > πi1, where πic represents the within-class item difficulty for item i, MA32, is the probability (median across all SMEs) of a correct response for examinees at the border of classes 3 (Advanced) and 2 (Proficient) for a particular item, and MA21, is the probability (median across all SMEs) of a correct response for examinees at the border of classes 2 (Proficient) and 1 (Developing) for the same item.

Results from the 3-class Angoff LCA were compared to the championed model from other standard setting techniques.

Estimation

Maximum likelihood (ML) estimation was used to estimate all models (i.e., general LCA, ordinal LCA, Angoff LCA) via the expectation maximization (EM) algorithm in Mplus Version 7.3 (Muthén & Muthén, 1998-2012). ML estimation is used to find the parameter values for which the data are most likely (Enders, 2005). ML estimation utilizes the log likelihood distribution, which captures how “likely” the sample data are for various sets of parameter values (Enders, 2005). The goal of ML estimation is to find the set of parameter values associated with the highest log likelihood value (Enders, 2005).

Typically, ML estimation centers on a global maximum, or the highest peak of the log likelihood distribution, regardless of the starting values used for the parameter estimates. However, an issue with ML estimation for mixture models is a tendency for the distribution to have multiple peaks (Masyn, 2013; Pastor & Gagné, 2013)—leading to local maxima in addition to a global maximum. Although the goal of ML is to arrive at the parameter estimates associated with the global maximum, a researcher might arrive at the parameter estimates associated with a local maximum (which are not the parameter
estimates for which the data are most likely). Issues with model estimation, such as incorrect parameter estimates, often occur when an estimation algorithm converges on a local likelihood maximum and not a global likelihood maximum (Geiser, 2012; Uebersax, 2000). As a model becomes more complex (i.e., the number of latent classes grows), the potential of converging on a local maxima increases (Uebersax, 2000).

To avoid such convergence issues, there are two stages involved in ML estimation using the EM algorithm in Mplus. In the initial stage, a variety of random starting values are used to find the solution with the best possible log likelihood value. The model is estimated a prespecified amount of times (defined by the user) with a different set of random starting values that are generated each time for the model parameters (Geiser, 2012). If only a limited number of starting values is used for a LCA model, there is an increased chance of encountering a local likelihood maximum and the estimation technique is likely to produce inaccurate parameter estimates (Geiser, 2012). Thus, a large number of random starting values is highly recommended in order to avoid model termination at a local maximum (Geiser, 2012). Following the advice of Geiser (2012), 1,000 random sets of starting values were used in the initial stage of the EM optimization process.

In the second stage of model estimation, a specific number (e.g., 100) of random starting values with the largest log likelihoods are selected from the results of the first step of the optimization process and used as the starting values for the final stage optimizations (Geiser, 2012). For example, the 100 starting values with the largest log likelihoods in the initial phase may be specified by the user as the starting values in the second step of the optimization process. Based on Geiser’s (2012) recommendations,
specifications were made for Mplus to select the 100 starting values with the largest log likelihoods in the initial phase to be used in the second step of the optimization process. Iterations of the 100 starting values were performed until convergence of the model was achieved, which is usually defined as a parameter change of less than 0.000001 in Mplus (Geiser, 2012). The final solution corresponded to the starting value associated with the highest log likelihood from this set.

**Model Fit**

A variety of fit indices were used to compare the models. The log-likelihood ($LL$) was obtained for each LCA model to evaluate the likelihood of the data, given the model parameters. Higher $LL$ values indicate superior fit of the model to the data. However, $LL$ values always increase as model complexity increases. To account for model complexity, information criteria (IC) were used as an additional source of evidence to assess model-data fit. The Bayesian information criterion (BIC; Schwarz, 1978), sample-size adjusted BIC (SSA-BIC; Sclove, 1987), and Schwarz information criterion (SIC; Schwarz, 1978) were used in the present study based on prior research that suggests these many of these indices function well for mixture modeling techniques (Henson, Reise, & Kim, 2007; Masyn, 2013; Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2008; Yang, 2006). The IC values were compared within and across the general, ordinal, and Angoff LCA models to champion a particular solution. The model associated with the smallest IC value was considered superior. The Lo-Mendell-Rubin likelihood ratio test (LMR; Lo, Mendell, & Rubin, 2001) and bootstrap likelihood ratio test (BLRT; McLachlan & Peel, 2000) were also used to evaluate whether a $k$ class model fit significantly better than a $k$-1 class model. $P$-values less than 0.05 for the LMR and BLRT indicate the more complex
model (i.e., with \( k \) classes) should be championed as a superior fit to the data. It is also important to note the LMR and BLRT can only be used to compare models with different numbers of classes that are of the same type. For instance, the 1-class and 2-class general LCA can be compared using the LMR and BLRT, but a 1-class general LCA cannot be compared to a 2-class ordinal LCA using these tests. Additionally, the LMR cannot be computed when constraints are placed upon the model in Mplus. Thus, the LMR can only be used to compare general LCA class solutions. Lastly, the Bayes Factor (BF) was used to compare the relative fit of models in the study. Just like the IC described earlier, the BF can be used to compare any two competing models (Masyn, 2013). To compute the BF, Equation 7 was used.

\[
BF_{A,B} = \exp[SIC_A - SIC_B]
\]

As \( BF_{A,B} \) grows larger in Equation 7, the evidence is increasingly in favor of Model A and less in favor of Model B (Masyn, 2013). If a BF value less than 1 is found using Equation 7, then the evidence shifts to be in favor of Model B (Jeffreys, 1961). The BF values were compared within and across the general, ordinal, and Angoff LCA models.

**Classification Accuracy**

The classification accuracy of models, which is based on posterior probabilities of class membership, was also considered in the current study. As aforementioned, the posterior probability of class membership can be estimated for every examinee in every class of an LCA model using Equation 4. If the posterior probability of membership within a given class is very likely or unlikely (i.e., near 1 or 0, respectively), there is great certainty in the classification of an examinee. Conversely, as the posterior probability of
membership within a given class strays from 1 or 0, there is greater skepticism about the resulting classification of that examinee.

To ascertain the classification accuracy of each model, the classification table and entropy statistic were consulted in Mplus. Calculation of the classification table is based on modal assignment of examinees to the class for which their posterior probability is largest. The classification table in Mplus presents the average posterior probability of membership for all examinees that were assigned to each class. For example, in a 2-class solution, the average posterior probabilities in Classes 1 and 2 are shown for examinees actually assigned to Classes 1 and 2. When classification accuracy is high, the Class 1 averages will be largest for examinees assigned to Class 1 and the Class 2 averages will be largest for those assigned to Class 2. The relative entropy statistic, which ranges from 0 to 1, also captures classification accuracy (Ramasway, DeSarbo, Reibstein, & Robinson, 1993). The relative entropy statistic is calculated using the posterior probabilities of class membership \( P(c|X) \), the sample size \( N \), and the number of classes \( C \) as follows (Masyn, 2013):

\[
\frac{\sum_{i=1}^{N} \sum_{c=1}^{C} \left[ -P(c|X) \ln(P(c|X)) \right]}{N \log(C)}
\]  

(8)

**Analyses Pertaining to Research Questions**

Five sets of RQs were examined in the current study. Each set of RQs will be individually addressed in the following section.

**Research Questions, Set 1: Comparison of General LCA and Ordinal LCA Models.** There are two research questions in Set 1: Which model is championed as the best fit to the data for the general LCA and the ordinal LCA? Do the general and ordinal LCA each indicate a consistent number of classes as the best solution? To answer the first
research question in this set, model-fit indices were compared across solutions within each model type. For example, the $k$ and $k-1$ class general LCA solutions were compared to one another. For these comparisons, the BIC, SSA-BIC, SIC, BF, LMR and BLRT were consulted. The second research question in this set was answered by considering whether the number of classes associated with the most favorable solution for the general LCA model was equal to or different from the number of classes associated with the most favorable solution for the ordinal LCA model. Recall, there is no reason to compare different solutions for the Angoff LCA because only a 3-class solution was fit to the data for that model.

**Research Questions, Set 2: Comparison of Championed General LCA, Ordinal LCA, and Angoff LCA Models.** There are two research questions in Set 2: How do the results from the championed general LCA model, championed ordinal LCA model, and Angoff LCA compare? Are the class-specific item difficulties similar or different? To answer this set of research questions, the within-class item difficulty values were compared for each of the three models.

**Research Questions, Set 3: Comparison of Model-Fit for Approaches Using Statistical Techniques.** A single research question was posed in Set 3: Across the three approaches involving statistical techniques, which model is championed as the best fit to the data? The model-fit indices that can be used for comparisons of models of different types were employed. The BIC, SSA-BIC, SIC, and BF were all considered to make a determination about which championed solution fit the data best.

**Research Questions, Set 4: Percentage Classification into Performance Groups across Standard Setting Methods.** Two research questions were posed in Set 4:
Are similar percentages of students classified into each performance group using the various approaches to standard setting (i.e., modified Angoff, general LCA, ordinal LCA, and Angoff LCA)? What do the ERIT-XA test score distributions look like within the groups produced by each standard setting approach, and how do those compare across approaches? For the modified Angoff, the percentage of students classified into the Developing performance group was determined by calculating the proportion of students below the first cut score from the modified Angoff. The percentage of students classified into the Proficient performance group was determined by calculating the proportion of students who scored at or above the first cut score but lower than the second cut score. The percentage of students classified into the Advanced performance group was determined by calculating the proportion of students who scored at or above the second cut score. For the standard setting approaches that used a statistical technique for classification, the percentage of examinees assigned to each performance group was calculated using the posterior probabilities of class membership for each championed solution.

**Research Questions, Set 5: Comparison of External Validity Evidence across Standard Setting Methods.** There are two research questions in Set 5: Do the relations between classifications and external variables depend on the standard setting method used? Also, do the relations with external variables align with expectations, thus providing validity evidence for the standard setting method used? The relation between group membership and attitudes toward ethical reasoning, amount of exposure or “dosage” to ethical reasoning, self-reported effort on tests from Assessment Day, and SAT Reading test scores were examined across the performance categories for the
championed solution from each standard setting approach. The details of the variables and analyses used to explore RQ5 are presented in the subsequent section on validity.

Validity

If differences truly exist between student levels of performance on the ethical reasoning knowledge and abilities measured by the ERIT-XA, one might expect individuals assigned to separate performance groups to differ on other variables of ethical reasoning. For example, the group scoring highest on the ERIT-XA might be anticipated to have greater exposure to the 8KQ or more interest in ethical reasoning compared to the other groups. To evaluate the validity of classifying students into performance groups on the ERIT-XA and to assess whether one method was associated with stronger validity evidence than another, the relation between class membership and other variables related to ethical reasoning was analyzed for each standard setting method. The external evidence obtained to validate the classification of examinees into performance groups is described below.

Auxiliary Variables.

Survey of Ethical Reasoning (SER). Just like the ERIT-XA, the SER is one of a battery of instruments administered by the MC to assess the ethical reasoning knowledge, skills, and abilities of students at JMU (Smith et al., 2015). Because the ERIT-XA and SER are both used to assess aspects of ethical reasoning, it seems reasonable to expect performance on the two assessments to be related. The SER is comprised of four sections, each of which will be discussed below.

The first section of the SER asks students to rank-order 10 different skills—such as critical thinking, writing, and ethical reasoning—from 1 (Most Important) to 10 (Least...
Important). The average rank of ethical reasoning as an important skill was examined across all performance categories of the ERIT-XA for each standard setting method. The second section of the SER presents five statements about perceived importance of ethical reasoning and five statements about confidence in applying the ethical reasoning process. Response options range from 1 (Strongly Disagree) to 5 (Strongly Agree) on a five-point Likert scale. Higher scores on the Importance subscale indicate students find ethical reasoning to be more important than students with lower scores. Likewise, higher scores on the Confidence subscale indicate students feel more confident about applying the ethical reasoning process than students with lower scores. The average score on the Importance and Confidence subscales were compared across all performance categories of the ERIT-XA for each standard setting method. The third section of the SER asks students to report how frequently they engage in various ethical reasoning behaviors using a five-point Likert scale ranging from 1 (Never) to 5 (Daily). Higher scores indicate students engage in ethical reasoning more often than students with lower scores. The average frequency of engaging in ethical reasoning behavior was compared across all performance categories of the ERIT-XA for each standard setting method. The fourth and final section of the SER is not described here and was not used for validity analysis because it is more relevant to the importance of the different Key Questions to students than the ethical reasoning behaviors and abilities they exude. Responses from the first three sections of the SER were compared across all performance categories of the ERIT-XA for each standard setting approach. Of the ERIT-XA sample used for analysis containing 3,558 students, 2,576 had complete data on the SER.
“Dosage” Data. Beginning in Spring 2015, five questions were appended to the ERIT-XA. These questions asked students how much exposure or “dosage” they have had to the 8KQ through various activities while at JMU. Specifically, the items asked students to self-report their level of engagement with the 8KQ through *It’s Complicated* (Dosage 1), activities outside of the classroom (Dosage 2), general education (Dosage 3) or major coursework (Dosage 4), and in the residence hall (Dosage 5). Response options on the dosage items range from 1 (*Never Taken*) to 5 (*Heavy Exposure*). Smith et al. (2015) showed ERIT-XA total scores were positively, statistically significantly correlated with the *It’s Complicated* and General Education dosage items (items 1 and 3, respectively). Thus, examining how students in different performance categories respond to the “dosage” items appears to be a useful method to evaluate the validity of performance group classifications.

The average response to dosage items 1, 3, 4, and 5 were compared across the classifications resulting from each standard setting approach. The wording of item 2 varied across the Spring 2015 and Spring 2016 administrations of the ERIT-XA, so the analyses only included four of the five items asked of students. Of the ERIT-XA sample used for analysis containing 3,558 students, 907 had complete data on the dosage items.

**Student Opinion Scale (SOS).** Many students complete the Student Opinion Scale (SOS; Sundre & Moore, 2002) as the final task at the end of their testing session during Assessment Day. The SOS is a self-report measure of examinee motivation that asks students about the level of effort they put forth during the testing session and the importance they assigned to their performance on tests taken (Sundre & Moore, 2002). The relation between SOS scores and performance on the ERIT-XA was compared across
the classifications resulting from the different standard setting approaches. Differences in SOS scores between performance groups may signify that examinee motivation, rather than the skills measured by the ERIT-XA, can explain the classification of students into discrete groups. Of the ERIT-XA sample used for analysis containing 3,558 students, 3,324 had complete data on the SOS items.

**SAT Reading Test.** Reading ability is another factor that may contribute to student performance on the ERIT-XA. The SAT Reading test provides an indication of a student’s aptitude across a range of reading skills, including identification of information, using words in context, and analyzing an experiment. Because the ERIT-XA requires students to read numerous scenarios and understand them in the context of an ethical dilemma, an examinee’s reading ability may be a contributing factor when making performance classifications using a test score. Students’ scores on the SAT critical reading test were compared across the classifications resulting from the standard setting approaches to evaluate whether higher scores on the ERIT-XA were a function of ethical reasoning skills or of higher reading ability. Of the ERIT-XA sample used for analysis containing 3,558 students, 2,916 had information regarding their performance on the SAT Reading test.

**Validity analyses.** A common approach to classify individuals into groups is to assign them to the class associated with their highest posterior probability and use statistical analyses to relate other variables to class membership. Analysis of Variance (ANOVA) and chi-square are common statistical tests used to examine whether groups statistically differ with regard to auxiliary variables used for validity evidence. ANOVA is often used if the auxiliary variables are continuous, whereas chi-square can be used if
the auxiliary variables are categorical. Given that all auxiliary variables (i.e., the SER and dosage totals) in the current study are continuous, ANOVA was used for the validity analyses for the three groups resulting from the modified Angoff and also for the classes emerging from the championed LCA model. The amount of variance explained by group membership in each auxiliary variable was obtained using omega-squared ($\omega^2$), which is a commonly reported effect size index for ANOVA. The amount of variance explained by group membership was used to ascertain if group differences on the auxiliary variables were more substantial for one approach over another.

An additional consideration for mixture models (e.g., LCA), however, is the extent to which examinees are accurately classified into performance groups. If classification accuracy is high, we have great certainty about the groups to which examinees are assigned. If classification accuracy is low, we are not confident about the groups to which examinees are assigned. Unfortunately, ANOVA, chi-square, and other common tests do not account for classification error in a model (Clark & Muthén, 2009). Consequently, the results of these tests may be misleading if classification accuracy is not high. The BCH procedure in Mplus, proposed by Bakk and Vermunt (2016), accounts for classification error of examinees in a mixture model and was recommended by Asparouhov and Muthén (2014) for use with mixture models in Mplus. The BCH procedure is a weighted ANOVA in which classification of individuals is based on their posterior probability of membership in each class and the ANOVA weights are based on the inverse of the classifications for individuals (Bakk & Vermunt, 2016; Vermunt, 2010). Thus, the BCH procedure was used to supplement the ANOVAs for the championed LCA model.
CHAPTER FOUR

Results

Results from the modified Angoff traditional standard setting will be reported first. The median rating of expected performance on each item is presented by borderline group, as are the cut scores established during each round of the standard setting and indicators of the variability in ratings by round. A summary of the final performance standards set by the panel and the percentage of students classified into each group if those standards were applied to a sample of students who took the ERIT-XA will also be presented. Findings from the data-driven and integrated approaches to standard setting are shared next, including a championed class-solution for the general LCA and ordinal LCA models. After presentation of the results for the general LCA, ordinal LCA, and Angoff LCA, each set of RQs is addressed.

Modified Angoff

Two rounds of ratings were conducted for the modified Angoff. After each round, SMEs were shown the individual ratings they provided for each item of the ERIT-XA as well as their individual raw and percent-correct cut scores for each borderline group. The raw score and percent-correct performance standards (i.e., cut scores) for each rater are shown by round in Table 5. The median performance standards were very similar across rounds. Round 2 ratings were used to establish the final performance standards on the ERIT-XA because those were the last ratings provided by SMEs. For the cut score separating Developing/Proficient performance, the final raw score standard was 30.48 (out of 50) and the final percent-correct standard was 60.95%. For the cut score separating Proficient/Advanced performance, the final raw score standard was 40.66 and
the final percent-correct standard was 81.31%. Table 5 also indicates there was a moderate amount of variability in the cut scores set by SMEs each round, although the variability in the ratings provided by the panel was reduced from Round 1 to Round 2—a good indication interrater consistency (i.e., reliability) was improved in Round 2. The feedback and discussion session following Round 1 ratings was designed to improve interrater reliability, so the diminished variability in ratings was expected.

The median of the panel’s ratings for each item of the ERIT-XA are shown by borderline group for Rounds 1 (Table 6) and 2 (Table 7). As expected, all median item ratings were substantially higher for the Minimally Advanced borderline group in both rounds, indicating SMEs believed those students were more likely to correctly answer each item.

Figure 10 visually conveys the consistency SMEs had across rounds when rating items that were perceived to be easy and the discrepancies for more difficult items. For instance, items 12 and 21 were rated as nearly the same (high) probability of correct response across rounds for each borderline group. Alternatively, the ratings for items 39 and 46 were inconsistent across rounds. For these two items, the median rating provided by SMEs dropped between 6.5 and 20 percentage points for each borderline group from Round 1 to Round 2. It is not necessarily a problem the median ratings were discrepant across rounds for some items—it is good SMEs were willing to modify their estimations of expected examinee performance after engaging in discussion/feedback—it merely appears SMEs were more lenient with their Round 1 ratings than Round 2. Overall, Figure 10 suggests item ratings were fairly similar across Rounds 1 and 2.
A few different forms of information about student performance on the ERIT-XA were provided to SMEs during the modified Angoff. Score distributions illustrating the performance of JMU students from four semesters on the ERIT-XA were provided to SMEs at the beginning of the standard setting. The distributions (Figures 11 and 12, respectively) showed performance in terms of raw scores and percent-correct on the ERIT-XA, with means and other summary statistics provided. Later, “impact” data demonstrating the percentage of students classified into each performance category using a subsample (n =1,840) of the larger ERIT-XA data set used for analyses were shared with the panel. Results showed 24% of students were classified as Developing, 55% as Proficient, and 21% as Advanced. Upon receiving impact data, the panel engaged in discussion about the suitability of the performance standards established after Round 2. Ultimately, the panel determined the standards were sufficient and chose not to modify any cut scores.

**General LCA**

Five general LCA solutions were fit to the data, ranging from 1 to 5 classes. Profile plots for the 1-, 2-, 3- and 5-class general LCA solutions can be found in the Appendix and the plot for the 4-class solution is presented in Figure 13, which will be discussed later in the results section. As shown in Table 8, the BIC, SSABIC, and SIC fit indices all moved closer to zero (an indicator of better model-data fit) as the number of classes increased—until the 5-class solution was estimated. The BIC and SIC were further away from zero for the 5-class solution than the 4-class solution and the LMR was not statistically significant for the 5-class solution. Additionally, the probability of a correct response to item 46 approached a boundary parameter (i.e., 0 or 1) in the 5-class
solution. Typically, the LCA model is rejected when item probabilities approach a boundary parameter (Geiser, 2013). Further, Class 5 was comprised of only 3.4% of students. Taking into account all results from the 5-class general LCA solution, there was convincing evidence the model should not be championed.

Therefore, the comparison of interest was whether the 3-class solution or 4-class solution was the best-fitting general LCA model to the data. The fit indices and statistical tests primarily favored the 4-class model over the 3-class model. The BIC, SSABIC, and SIC were all closer to zero for the 4-class solution, and the LMR and BLRT were both statistically significant. Only the BF indicated the 3-class solution fit the data better than the 4-class solution. As a result, the 4-class solution was championed as the best-fitting general LCA model.

**Ordinal LCA**

Only two ordinal LCA solutions were fit to the data, the 4-class and 5-class solutions. A 1-class ordinal LCA solution was not analyzed because there is no need to order groups in such a model; there is only one group. The 2- and 3-class general LCA solutions both produced perfect ordering of item difficulties across classes. That is, for the 2-class general LCA solution the item difficulty was higher for Class 1 than Class 2 on all items. Likewise, the item difficulties were higher for Class 1 than Class 2 on all items in the 3-class general LCA solution, and the item difficulties were also higher for Class 2 than Class 3 on all items. Thus, the 1-, 2-, and 3-class ordinal LCA solutions were not fit to the data because they are identical to the corresponding general LCA solutions. The model-fit indices (e.g. LL, BIC, SSABIC) for the 2- and 3-class general LCA solutions equal those for the 2- and 3-class ordinal LCA solutions, respectively, because
the solutions are equivalent. Profile plots for the 1-, 2-, and 3-class ordinal LCA solutions are not shown because they were identical to the general LCA solutions (which are available in the Appendix).

Three constraints were made to parameters in the 4-class general LCA solution to produce the 4-class ordinal LCA solution. Class 3 was constrained to have a higher probability of correct response than Class 4 on items 2 and 39, and Class 2 was constrained to have a higher probability of correct response than Class 3 on item 46. The profile plot for the 4-class ordinal solution is presented in Figure 14, which will be discussed later in the results section. A number of problems were again encountered when estimating the 5-class solution and the model failed to converge. The lack of convergence was likely a result of estimated item difficulties approaching the boundary parameter in Class 1. Thus, no results are reported for the 5-class ordinal LCA solution.

Therefore, the comparison of interest was whether the 3-class solution or 4-class solution was the best-fitting ordinal LCA model to the data. The BIC, SSABIC, and SIC were all closer to zero for the 4-class solution. Although the LMR statistic cannot be computed when making model constraints in Mplus, the BLRT was statistically significant for the 4-class ordinal LCA solution—indicating the model fit better than the 3-class ordinal LCA solution. Only the BF suggested the 3-class solution fit the data better than the 4-class solution. Based on the model-fit indices and statistical tests, the 4-class solution was championed as the best-fitting ordinal LCA model.

Angoff LCA

The Angoff LCA 3-class solution was tested a number of different ways, none of which were able to achieve convergence. First, parameter constraints were placed on the
model for all items using the ratings from SMEs as part of the modified Angoff standard setting. When that method failed, constraints were only placed on items that did not meet the thresholds established by SMEs during the modified Angoff, which was in line with the recommendations of Croon (1990; 2002) and also done for the ordinal LCA. A final attempt was made by providing starting values for each parameter of the Angoff LCA that was estimated as part of the analyses in Mplus. Despite these efforts, the model never reached convergence.

**Research Questions**

The remaining paragraphs of this chapter focus on results as they relate to the RQs that guided this study.

**RQ1**: Which model is championed as the best fit to the data for the general LCA and the ordinal LCA? Do the general and ordinal LCA each indicate a consistent number of classes as the best solution? To address RQ1, the championed solution for the general LCA and ordinal LCA models were selected using model-fit indices and compared for consistency in number of classes. The 4-class solution was championed for both the general and ordinal LCA models, so there was consistency in the number of classes for the two championed data-driven models.

**RQ2**: How do the results from the championed general LCA model, championed ordinal LCA model, and Angoff LCA compare? Are the class-specific item difficulties similar or different? To address RQ2, item difficulties for only the championed general LCA and ordinal LCA solutions (i.e., the data-driven standard setting techniques) were compared because the Angoff LCA model did not converge. As previously mentioned, profile plots were created for the championed 4-class general and
ordinal LCA solutions to illustrate the within-class item difficulty for each model. Figure 13 displays estimates of the probability of correct response by item of the ERIT-XA for the 4-class general LCA solution. The ultimate takeaway from the 4-class general LCA profile plot is that groups were ordered from low to high on nearly all items of the ERIT-XA, which has implications for the how the championed general LCA and ordinal LCA solutions compare. Figure 14 displays estimates of the probability of correct response by item of the ERIT-XA for the championed 4-class ordinal LCA solution. Indeed, the profile plots for the two championed LCA models were nearly identical, which makes sense given only three parameters needed to be constrained from the 4-class general LCA to produce the 4-class ordinal LCA. The only slight differences were the ordering of two groups on items 2, 39, and 46. Interestingly, items 2, 39, and 46 were three of the most difficult items on the ERIT-XA for students in the current sample (the 1-class general LCA solution in the Appendix visually conveys item difficulties for the overall sample).

**RQ3: Across the three approaches involving statistical techniques, which model is championed as the best fit to the data?** To answer RQ3, only the championed models from each data-driven standard setting approach were compared using appropriate model-fit indices because the Angoff LCA failed to converge. The championed general and ordinal LCA solutions produced nearly identical model-fit results. Although the differences between the fit indices were minute for the two models, the BIC, SSABIC, and SIC were all very slightly in favor of the 4-class general LCA solution (Table 8). The BF also provided evidence—albeit weak, according to Jeffreys’ (1961) scale of evidence for Bayes factors—the 4-class general LCA solution fit the data best (BF = 1.10). However, as alluded to previously, the classes were unordered for three
items in the 4-class general LCA. The intention of standard setting is to classify students into ordered groups on all items. Because the 4-class ordinal LCA solution was more aligned with the purpose of standard setting than the 4-class general LCA solution, and the fit of the two models was practically equivalent, the 4-class ordinal LCA was championed as the best-fitting data-driven model.

**RQ4: Are similar percentages of students classified into each performance group using the various approaches to standard setting (i.e., modified Angoff, general LCA, ordinal LCA)?** What do the ERIT-XA test score distributions look like within the groups produced by each standard setting approach, and how do those compare across approaches? To answer the two questions for RQ4, a few analyses were performed. The size of each performance group that resulted from the modified Angoff is reported first for the entire sample of students in the study. An analysis regarding the percentage of students assigned to each class using the data-driven approaches (i.e., general LCA and ordinal LCA) follows. Attention is then turned to the extent to which results from the traditional and data-driven standard setting methods correspond. Test performance for students in the groups resulting from the two different standard setting approaches is described and the level of agreement between the two is considered.

For the traditional standard setting, the performance classification analyses were rerun using the entire ERIT-XA data set \(N = 3,558\) after the modified Angoff was completed. Results revealed 26.08\% of students were classified as *Developing* (i.e., scored below 30.48 on the ERIT-XA), 54.30\% of students were classified as *Proficient* (i.e., scored between 30.48 and 40.66 on the ERIT-XA), and 19.62\% of students were
classified as *Advanced* (i.e., scored at 40.66 or higher on the ERIT-XA). As shown in Table 9, the mean score on the ERIT-XA was lower for students classified as *Developing* ($M = 24.28; SD = 6.05$) than those classified as *Proficient* ($M = 35.64; SD = 2.80$). The means for students classified as *Developing* or *Proficient* were both lower than the mean for students classified as *Advanced* ($M = 43.12; SD = 1.85$). The average score on the ERIT-XA including all students in the sample was 34.14 (68.28% correct; $SD = 7.54$). The typical, or average, student in the sample would be classified as *Proficient* in their ethical reasoning abilities, as measured by the ERIT-XA, using the cut scores derived from the modified Angoff standard setting.

The percentage of students assigned to each class in the championed general LCA and ordinal LCA models was based on modal assignment of examinees to the class for which their posterior probability was largest. In the general LCA, Class 1 constituted about 33% of the sample and was generally characterized by a high likelihood of correct response across all items of the ERIT-XA. Class 2 encompassed nearly half of the students in the sample (48.4%) and was the largest performance group. Students in Class 2 generally did well on the ERIT-XA but were not as consistent in their performance as students in Class 1; they were more variable in their likelihood to correctly answer each item than students in Class 1. Class 3, which contained 15.0% of students, had lower probabilities of correct response across nearly all items of the ERIT-XA compared to Class 2. The lowest achieving group in the 4-class general LCA model was very small (3.4%) and had a probability of less than 0.50, on average, of correctly answering any item. All class sizes and item response profiles from the championed general LCA solution remained almost exactly the same in the championed ordinal LCA solution.
Class 1 comprised about 33% of students in the sample, Class 2 encompassed nearly half of students (47%), Class 3 contained approximately 16% of students, and Class 4 included less than 4% of the sample.

Entropy, which is a measure of classification accuracy, was equivalent and high (Entropy = 0.815) for the championed general and ordinal LCA models. Table 10 presents the average posterior probabilities of class membership by latent class of the championed ordinal LCA solution. Rows include students who were assigned to that particular latent class and columns represent the average posterior probability of class membership by all latent classes for students assigned to a particular class. Values along the diagonal of Table 10 indicate the average posterior probability of students being in the class to which they were assigned (based on the estimated model). All probabilities along the diagonal are greater than 0.872, which reveals students had a high probability of being assigned to their most likely latent class in the championed ordinal LCA solution. Because entropy was high in the championed general and ordinal LCA solutions, percent classifications using either modal assignment or the estimated parameters in the model were very similar. Percent classifications based on the estimated parameters in the 4-class general and ordinal LCA models are reported in Figures 13 and 14, respectively.

The data-driven standard setting approaches indicated students should be classified into a different number of performance groups than was used for the traditional standard setting. Three groups were used for the modified Angoff, whereas both the general LCA and ordinal LCA championed solutions suggested four groups was optimal for classification purposes. Test score statistics for each group/class of the modified
Angoff and 4-class ordinal LCA—which was the championed data-driven model—are presented in Table 9. Based on test scores alone, the average student in Class 1 would be classified as Advanced, in Class 2 as Proficient, and in Classes 3 and 4 as Developing. Although it generally appears there was correspondence between a particular class of the ordinal LCA and a performance category of the modified Angoff when considering average test score performance in each class, that was not the case for all students within a class due to the within-class variability in test scores. In other words, not all students from each class would be classified into a single performance group if the modified Angoff cut scores were used instead. For instance, the minimum values for Classes 1 and 2 extend below the cut scores for the Advanced and Proficient performance groups, respectively. Likewise, the maximum value for Class 3 extends beyond the cut score for Developing performance. All students in Class 4, however, did have a total ERIT-XA score within range of the boundaries for the Developing performance group; that is, the minimum (3) and maximum (21) observed test score for Class 4 was beneath the cut score for the Proficient performance.

An analysis was conducted to investigate the level of agreement in performance classification between the modified Angoff and 4-class ordinal LCA solution. Classifications based on cut scores from the modified Angoff procedure were compared to those for the ordinal LCA analysis, in which modal assignment of students to classes was based on a student’s largest posterior probability of class membership. Full results of the classification analysis for the two methods are presented in Table 11. An initial look examined the breakdown of class membership for students in the 4-class ordinal LCA by performance category of the modified Angoff. Almost all students classified as Advanced
using the modified Angoff were members of Class 1 (i.e., the highest performing group on the ERIT-X). Only four students in the *Advanced* performance group were members of Class 2, and none were assigned to Classes 3 or 4 (i.e., the lower performing groups on the ERIT-XA). There was a greater mix of class membership for the *Proficient* performance group. Nearly 75% of students who were classified as *Proficient* were assigned to Class 2, but about 25% of *Proficient* students were assigned to Class 1. Very few *Proficient* students were assigned to Class 3 and zero to Class 4. Even more disagreement was found between the classification methods for the *Developing* performance category. Less than 60% of students in the *Developing* group were in Class 3. The remaining percentage of *Developing* students was largely assigned to Class 2, and roughly another 13% were assigned to Class 4; none were members of Class 1.

Table 1 can also be summarized by considering the percentage of students in each ordinal LCA class who were grouped into the various modified Angoff performance categories. About 60% of students in Class 1 were categorized as *Advanced*, although another 40% were classified as *Proficient*. No students in Class 1 were classified as *Developing*. A large majority of students in Class 2 (about 83%) were grouped into the *Proficient* performance category. Approximately 16% of students in Class 2 were classified as *Developing* and only four students were in the *Advanced* performance group. Nearly all students from Class 3 fit neatly into the *Developing* performance group, while eight others were in the *Proficient* group. Complete agreement between the two standard setting methods was found for Class 4; all students in Class 4 were categorized as *Developing* using the modified Angoff.
To further consider the level of association between the methods, a chi-square test of independence was conducted to test the null hypothesis that there was no relation between performance group classification for the modified Angoff and class membership in the 4-class ordinal LCA. Results indicated there was a statistically significant relation between performance group classification and class membership, \( \chi^2 (6, N = 3,558) = 3,771.03, p < .0001 \). Overall, there appears to be general agreement in performance classifications made using the traditional and data-driven approaches, but the agreement is not perfect. Further discussion of the similarities and differences between classifications made using the modified Angoff and ordinal LCA will be explored in Chapter Five.

**RQ5: Do the relations between classifications and external variables depend on the standard setting method used?** Also, do the relations with external variables align with expectations, thus providing validity evidence for the standard setting method used? Validity analyses were conducted using several variables for the modified Angoff and the 4-class ordinal LCA solution—the championed method for the data-driven approach. A summary of the average scores, variability, and sample size for each auxiliary variable is provided in Table 12.

Detailed results of the validity analyses for both standard setting approaches can be found in Table 13. For the modified Angoff, a one-way ANOVA was conducted to determine whether groups differed in their scores on the auxiliary variables and Tukey’s post hoc test was used for all pairwise comparisons. Statistically significant differences were found across groups for a number of variables, including scores on the SER Importance, Confidence, and Engagement subscales, Dosage item 1, SOS importance
scores, SOS effort scores, and SAT Reading test scores. On all auxiliary variables, groups were ordered such that the highest average score belonged to the Advanced performance group and lowest to the Developing group. For example, average scores on the SER subscales were highest for the Advanced performance group and lowest for the Developing group. Likewise, the Advanced group performed best on the SAT Reading test and the Developing group performed worst. A rule of thumb for omega-squared ($\omega^2$) is that values of .01, .06, and .14 indicate a small, medium, and large effect, respectively (Cohen, 1988). A small effect was found for the mean differences between modified Angoff performance groups on the SER Importance, Confidence, and Engagement subscales, responses to Dosage item 1, and SOS effort scores. A large effect was found for the mean differences between performance groups on SAT Reading scores. Omega-squared indicated there was no effect for the differences between performance groups on SOS importance scores.

For the 4-class ordinal LCA, a one-way ANOVA was conducted using the BCH procedure in Mplus to account for classification error of examinees. Traditional one-way ANOVA was also conducted for each of the ordinal LCA validity analyses in order to compute effect sizes, which cannot be calculated using solely the BCH procedure in Mplus. Statistically significant differences were found across groups for a number of variables, including scores on all SER subscales, Dosage items 1 and 3, SOS importance scores, SOS effort scores, and SAT Reading test scores. Again, groups were ordered on all auxiliary variables such that the Advanced performance group had the highest average scores and the Developing group had the lowest. Statistical significance results from the one-way ANOVA were aligned with those using the BCH procedure, which was not
surprising given the high entropy of the championed 4-class ordinal LCA solution. A small effect was found for the mean differences between ordinal LCA classes on the SER Importance, Confidence, and Engagement subscales, responses to Dosage item 1, and responses to Dosage item 3. A medium effect was found for the mean differences between classes on SOS effort scores. A large effect was found for the mean differences between classes on SAT Reading scores. Omega-squared indicated there was no effect for the differences between classes on the SER Rank subscale or SOS importance scores. Implications of the findings from the validity analyses will be considered in Chapter Five.
CHAPTER FIVE

Discussion

The purpose of this study was to describe and illustrate the various approaches to standard setting. Specifically, traditional, data-driven, and integrated approaches to standard setting were implemented and the results from each approach were evaluated and compared. Unfortunately, the model used for the integrated approach did not achieve convergence and results were not obtained. Reasons for the lack of convergence will be entertained in this chapter. The two models applied as data-driven classification techniques did converge, however. Many similarities were found between the two data-driven models that will be explored throughout this chapter, and those findings will also be compared to results from the traditional approach. Following an appraisal of the results for the traditional and data-driven approaches, a recommended method will be proposed for the ERIT-XA standard setting. Considerations for the three approaches to standard setting will be summarized near the end of the chapter and suggestions for future research will be made. A final section will be devoted to overall conclusions drawn from the study.

Nonconvergence of the Integrated Approach

The argument for an integrated approach to standard setting is that it offers the advantages of both the traditional and data-driven approaches, while also overcoming some of their downfalls. An integrated approach seems like the best option for setting standards because it allows researchers to classify examinees into performance groups by combining the judgment of SMEs with empirical data. However, integrated approaches proposed thus far have been difficult to implement because they are not synchronized
with commonly-used statistical software, require knowledge of Bayesian data analysis methodology, and have only been used with examinee-centered traditional standard setting methods. The integrated approach proposed in the current study (i.e., the Angoff LCA) attempted to address those problems by using a test-centered method that was compatible with common statistical software and did not rely on the use of Bayesian methods.

Although the Angoff LCA appeared to be a promising approach to standard setting, unfortunately the model never converged even though it was tested in multiple ways. Failure to achieve convergence can occur for a number of reasons in mixture models, including the estimation of a model that is inappropriate for the data or specifying a model that contains an incorrect number of latent classes (Masyn, 2013; Muthén & Muthén, 1998-2012). Thus, it is possible the lack of convergence for the Angoff LCA may have been due to the many constraints put on the model, rendering it inappropriate for the data. The two primary constraints made in the Angoff LCA included: (1) the number of classes, which was constrained to three to align with the performance categories in Table 3, and (2) the item difficulties, which were constrained to align with SME ratings. Specifically, constraints were placed on item difficulties in the Angoff LCA that did not meet the suggested probability of correct response provided by SMEs during the modified Angoff in the 3-class general LCA. Of the 150 item parameters estimated for the Angoff LCA, constraints were placed upon 56 parameters. If lack of convergence was indeed due to the model being inappropriate, it is worthwhile to speculate as to why it was inappropriate. Perhaps the ratings provided by SMEs did not align well with the data (i.e., with how students are performing) or the number of classes
that was specified was incorrect. A final possibility is that the ERIT-XA performance category descriptions crafted for the modified Angoff standard setting, and on which the SME ratings were based, may have been inappropriate. If so, nonconvergence may have resulted from an incorrect specification of the KSAs needed to perform well on the ERIT-XA, which would have also affected the appropriateness of the Angoff LCA for the data.

The lack of convergence is not being used here to argue there are flaws with the modified Angoff ratings or performance category descriptions in Table 3, but rather to point out the questions raised when either an integrated standard setting model does not converge (as in this study) or does not yield favorable model-data fit. Templin and Jiao (2012) point this out as well—they repeatedly mention model-data fit may be an issue with integrated approaches and encourage more research and guidance on the issue.

**Performance of the Data-driven Approaches**

Given the Angoff LCA model did not converge, the only model-based approaches to standard setting considered were the general and ordinal LCAs. Before comparing the championed general and ordinal LCA solutions, the 2-class and 3-class solutions will be briefly discussed for each LCA model to provide context to the results. There was no difference in fit for the 2-class and 3-class general and ordinal LCA solutions because all groups were perfectly ordered in those models. If either the 2-class or 3-class solution fit the data best, crowning either the general LCA or ordinal LCA as the overall best-fitting model would have been inconsequential because the two models were equivalent.

However, neither the 2-class nor 3-class LCA solutions were championed as the best-fitting general or ordinal LCA models. Instead, the 4-class solutions were championed for each model. The number of classes indicated by the championed general
LCA and ordinal LCA models was consistent, likely due in large part to nearly-perfect ordering of classes across items of the ERIT-XA in the general LCA solution. All but three of the 50 items on the ERIT-XA were perfectly ordered (in terms of item difficulty) across all four classes of the championed general LCA solution. The only slight differences were the ordering of two groups on items 2, 39, and 46. Even so, the selection of a championed data-driven model was not straightforward. The model-fit indices for the 4-class general LCA and ordinal LCA were practically identical (Table 9), although they did very slightly favor the general LCA. Conversely, the ordinal LCA was better aligned with the purpose of standard setting because all groups were ordered on each item.

The critical question, then, is whether the added constraints on the three unordered items from the general LCA are worth it in the ordinal LCA. An argument can be made for either solution. Strictly based on model-fit and fewer parameter constraints, the general LCA should be championed, whereas greater alignment with the intent of standard setting leads to the ordinal LCA as the championed solution. There does not appear to be an easy or absolute answer in this situation. The decision for this study was based on the premise that unordered groups on any items detracted from the purpose of the standard setting. Further, only three constraints were placed on parameters for the 4-class ordinal LCA and those constraints forced the item difficulties to differ very little from the values that were freely estimated in the 4-class general LCA (between 1-2% in terms of probability of correct response between adjacent classes, which is also apparent in Figures 13 and 14). Thus, the 4-class ordinal LCA solution was championed as the overall best-fitting model. Nonetheless, a great deal of subjective judgment and
interpretation may be required on the part of the researcher when considering all factors in the decision-making process of which solution to champion.

**Connection between Ordered Classes and Factor Model**

The finding of ordered classes, while certainly beneficial and aligned with our intention to yield groups that differ quantitatively, should not be entirely surprising. The nearly perfect ordering of groups across all general LCA solutions that converged provides support—beyond evidence already collected by Smith et al. (2015; 2016)—that the ERIT-XA is a unidimensional scale. To understand why it is not surprising for a unidimensional test to yield ordered classes in LCA, the connection between LCA and a factor model (or IRT model) must be made explicit. Path diagrams shown in Figure 15 depict a LCA model in A and one-factor model in B. The primary difference between the two models in Figure 15 is that a LCA model explains the relations among items using a latent categorical variable (e.g., latent classes), whereas a factor model explains the relations among items using a latent continuous variable (e.g., latent factor score or theta). The path diagrams convey that both models use latent variables, but of different types, to explain relations among items.

Although the connection between LCA and a factor model has been made clearer, further explanation is still needed to clarify why it is not surprising to find ordered classes when a unidimensional test is used. To more fully address this issue, consider the profile plots shown for the 4-class ordinal LCA (Figure 14) and a one-factor model at three different levels of the factor (Figure 16) that were fit to the same data. As can be seen in Figure 16, response profiles are ordered for all examinees. Response profiles follow the same pattern, whether below, at, or above the factor mean, and only differ in elevation. In
fact, a 1-factor model can be thought of as a 1-class LCA in which within-class variability is permitted and a function of an examinee’s level (or ability) on the factor. A LCA with ordered classes that demonstrates good fit to the data suggests the simpler factor model—which estimates far fewer parameters—is appropriate for the data. Likewise, a factor model that demonstrates good fit to the data suggests an ordinal LCA might be plausible.

In the current study, finding ordered groups using a general LCA indicates students differ quantitatively on the ethical reasoning skills measured by the ERIT-XA rather than qualitatively (i.e., as demonstrated in Figure 5). Again, this is because groupings of students from the current sample were perfectly ordered in the 2- and 3-class general LCA solutions, and nearly ordinal in the 4-class general LCA solution. In light of the ordering of groups and evidence demonstrating unidimensionality of the ERIT-XA, the fit of the 1-factor model was compared to the championed general and ordinal LCA solutions. All model-fit indices favored the 1-factor model over the 4-class general LCA and 4-class ordinal LCA (Table 8). In fact, the BF indicated there was strong evidence in favor of the 1-factor model over the championed 4-class ordinal LCA (BF was greater than 3.59x10^{139}).

The favorable fit of the 1-factor model over the championed LCA solution prompts the question of whether performance standards and groups are needed for the ERIT-XA. Is how students score on the test simply enough to differentiate amongst student ability? Might the grouping of students by standard setting methods be categorizing a continuous variable, which ultimately results in a loss of powerful information (MacCallum, Zhang, Preacher, & Rucker, 2002)? Statistically speaking, the
answer to both questions is likely yes. However, from a practical point of view, there are advantages to classifying examinee performance into groups—especially if groupings are aligned with the performance category descriptions created for the ERIT-XA (Table 3). Performance standards and the resulting groups provide additional information that is often more easily interpretable than just a total score. Performance levels and labels deliver more meaning to test scores, can be used to establish a benchmark for performance, and can also aid in comparison of scores, so long as they are accurate and valid descriptors of the KSAs exhibited by examinees.

Comparing Traditional and Data-driven Approaches

Because the categorization of continuous scores is considered beneficial for the ERIT-XA, it is worthwhile to consider how categorizations based on the traditional approach compare to the data-driven approach. In the current study, students were classified into three performance categories using the modified Angoff and four classes in the championed ordinal LCA solution. Because the 4-class ordinal solution was championed as the best-fitting data-driven model and convergence was not obtained for the Angoff LCA, comparisons between the modified Angoff and 4-class ordinal LCA classifications were of primary interest.

A key similarity between the modified Angoff and championed ordinal LCA solution is that each produced ordered groups, which is a foundational component of standard setting. The modified Angoff must yield ordered groups by design; in contrast, there is no guarantee ordered groups will emerge from a LCA. Because the groups that emerged from the championed data-driven method were ordered (i.e., the ordinal LCA), the two approaches were similar in this regard.
The primary difference between the two approaches was in the number of groups. Students were classified into three performance groups using the modified Angoff and four classes in the championed ordinal LCA solution. There was an association in the classification of examinees into groups between the two approaches (Table 11) and a chi-square test revealed a statistically significant relation existed between classifications made using the two methods. Specifically, Class 1 of the championed ordinal LCA solution generally aligned with the Advanced performance category, Class 2 of the championed ordinal LCA solution generally aligned with the Proficient performance category, and Classes 3 and 4 both had a substantial percentage of students who were categorized as Developing. Although there was some alignment between classifications made using the two approaches, there was not a one-to-one correspondence.

Validity Results

Further analyses were needed to choose between the two standard setting approaches for the ERIT-XA because groupings from the modified Angoff and championed ordinal LCA solution were not the same. Analyses were conducted using auxiliary variables to provide validity evidence for the classifications resulting from each approach. The results were compared and contrasted to determine whether the evidence provided greater support for one approach.

Before comparing validity results across approaches, consideration is first given to whether the validity evidence was supportive of the groupings. In the modified Angoff, statistical significance was found for all SER subscales, Dosage item 1, SOS importance and effort scores, and SAT Reading scores. For the 4-class ordinal LCA, statistical significance was found for two additional auxiliary variables: the SER Rank subscale and
Dosage item 3. The effect sizes were fairly consistent across the standard setting methods. Most of the differences between group means within approaches were judged to be small (based on $\omega^2$), with the exception of SOS Effort scores in the championed ordinal LCA solution ($\omega^2 = 0.08$) and SAT Reading scores for both approaches ($\omega^2 = 0.18$).

Although specific hypotheses were not stated for the ordering of groups on the external variables, results of the validity analyses indicated higher-performing groups also had higher means on variables thought to be positively related to ERIT-XA skills for each method. For instance, students in the upper achievement group on the ERIT-XA (i.e., Class 1 in the ordinal LCA and the Advanced group in the modified Angoff) also scored highest on attitudes toward ethical reasoning (i.e., the various SER subscales) for both approaches. Students in the lower achievement group on the ERIT-XA (i.e., Class 4 in the ordinal LCA and the Developing group in the modified Angoff) also scored lowest on attitudes toward ethical reasoning.

Significant differences among groups that did not yield strong supportive validity evidence for either approach included those pertaining to SOS importance and SOS effort scores. These variables are not thought to be positively related to ERIT-XA skills yet higher-performing groups had higher means on them. Differing levels of effort indicate classifications into groups may not be primarily based on ethical reasoning KSAs, but rather are related to how hard a student tried on the test. Classifications based on effort rather than ethical reasoning ability pose a problem because they are misaligned with the performance category descriptions crafted for the modified Angoff.
SAT Reading scores indicated students in more advanced performance categories/classes (Class 1 and the Advanced group) performed better on the SAT than those in lower-achieving groups (Classes 3/4 and the Developing group). However, validity results for the SAT Reading test could be interpreted two ways. The link between reading ability and examinee classification may suggest classifications are a function of something other than ethical reasoning ability. If so, the discovery that classifications were related to other variables is problematic because it indicates test scores are a consequence of unintended factors (that contribute to construct-irrelevant variance in the scores) and does not align with the purpose of the ERIT-XA standard setting. On the other hand, a positive association between SAT Reading test scores and group classification may be expected if the SAT Reading test aligns well with the KSAs measured by the ERIT-XA.

As a reminder, the SAT Reading test measures a range of reading and critical thinking skills, such as locating evidence that leads to a reasonable conclusion, identifying how evidence is used to support claims, examining hypotheses, interpreting data, and considering the implications of results. Many of these skills seem to be in alignment with the evaluative criteria crafted for the three performance categories of the modified Angoff. There were two dimensions specified as part of the evaluative criteria SMEs used to guide their ratings of expected examinee performance on the ERIT-XA. The first dimension created to distinguish between the ERIT-XA performance categories pertained to knowledge and understanding of the 8KQs. A consistent level of knowledge and understanding as well as an ability to extrapolate beyond the description of the 8KQs provided on the test represented Advanced performance, whereas an uneven
understanding of the 8KQs and difficulty inferring beyond descriptions on the test characterized *Developing* performance. The capability of students to examine an ethical scenario, locate pertinent details that will help lead to a reasonable conclusion, and consider the ramifications of that conclusion seem to be captured by both SAT Reading scores and ERIT-XA scores. The second dimension revolved around a student’s ability to identify which KQ was most consistent with an ethical decision/rationale. A similar component was also measured by the SAT Reading test: identifying how evidence is used to support claims. The two dimensions of examinee ability used for the performance category descriptions of the ERIT-XA and the content on the SAT Reading appear to both tap into critical reading and critical thinking skills. If that indeed is the case, it should be no surprise the performance groups and classes resulting from the standard setting approaches performed differentially on the SAT Reading test. Ultimately, a final conclusion regarding the validity results for the SAT Reading test is complicated and should not be made solely in the context of this study. For now, either conclusion is plausible and the implications of either determination should be considered.

Recall, 242 of the 3,558 students included in the sample used for the current study were participants in a semester-long ethical reasoning intervention associated with Madison Collaborative (MC) during the Fall 2016 semester. A comparison of the classifications for students who did and did not participate in the MC intervention may shed additional light on the relation between exposure to ethical reasoning and performance on the ERIT-XA. Two chi-square tests of independence were conducted to examine the relation between participation in the MC intervention and performance classification using the modified Angoff and 4-class ordinal LCA. Results revealed there
was a statistically significant relation between participation in the MC intervention and performance group classification using the modified Angoff, \( \chi^2 (2, N = 3,558) = 11.85, p = .003 \). Standardized residuals indicated more students than expected were classified as Advanced in the MC subsample if participation in the ethical reasoning intervention was unrelated to modified Angoff performance group. Table 14 shows nearly 28.10% of students were classified as Advanced in the subsample that participated in the MC intervention, whereas only 19.00% of students who did not participate in the MC intervention were classified as Advanced—a difference of nearly 10 percent between subsamples. The second chi-square test revealed there was also a statistically significant relation between participation in the MC intervention and class membership in the 4-class ordinal LCA, \( \chi^2 (3, N = 3,558) = 19.05, p < .0001 \). Standardized residuals indicated more students were assigned to Class 1 (highest-achieving group on the ERIT-XA) and less students assigned to Class 2 (second-highest achieving group on the ERIT-XA) than expected if participation in the ethical reasoning intervention was unrelated to class membership in the 4-class ordinal LCA. Table 15 shows almost half (45.45%) of the students that participated in the MC intervention were assigned to Class 1, whereas only 32.12% of students who did not participate in the MC intervention were assigned to Class 1. Additionally, 37.60% of students in the MC subsample were assigned to Class 2 compared to nearly half (49.13%) of the non-MC subsample.

To summarize, the results from each standard setting approach were consistent: a greater percentage of students in the MC subsample were categorized into either the Advanced performance group in the modified Angoff or Class 1 in the 4-class ordinal LCA than for the non-MC subsample. The increased proportion of students classified into
the upper-most achievement group in the MC subsample for each standard setting method, compared to the non-MC subsample, suggests students who participated in the MC ethical reasoning curriculum and activities also performed better on the ERIT-XA than those who do not. The findings are in alignment with expectations and indicate participation in MC interventions are fruitful for improving students’ ethical reasoning skills at the foundational level (i.e., for those skills measured by the ERIT-XA).

Overall, there was evidence the relations between classifications and scores on external variables did not differ by standard setting approach. When comparing the pairwise tests and mean values across approaches, similar results were found. Any differences in validity results that did emerge between approaches can be attributed to the presence of Class 4 and its substantially lower means on many of the auxiliary variables, relative to the other classes. However, because the presence of Class 4 leads to more statistically significant differences and larger differences among groups, it could be argued the validity evidence supports the ordinal 4-class solution over the modified Angoff. Indeed, the lowest-performing group in the 4-class ordinal LCA solution (i.e., Class 4) is also lower on variables thought to be related to the ERIT-XA learning objectives, which suggests students in this class are meaningfully different from students in other classes and should be retained.

The question becomes: How relevant is Class 4 to the ERIT-XA standard setting? Results showed all students assigned to Class 4 were also categorized as Developing using the modified Angoff cut scores. Does retention of Class 4 and use of the championed ordinal LCA solution help us make more informed decisions regarding students’ achievement on the objectives measured by the ERIT-XA? The short answer:
Probably not. Although Class 4 had lower means on the ERIT-XA and variables related to the ERIT-XA (e.g., SER subscales), their self-reported effort on the test was much lower than the other classes. It is quite possible students in Class 4 did not try on the test (or the SER, which was administered during the same testing session as the ERIT-XA) and did not produce scores indicative of their true level of ethical reasoning ability, as measured by the ERIT-XA. Given all the validity evidence regarding Class 4 of the championed ordinal LCA solution, it seems, at the very least, that students in Class 4 should be filtered from the data set because their scores on the ERIT-XA are not trustworthy.

It is worth mentioning the use of LCA was beneficial to this study because it helped uncover problems associated with Class 4. However, LCA also showed that classification of examinees using a data-driven approach may result in a class or group that is not a function of the KSAs being measured by a test (e.g., Class 4). Thus, the collection of validity evidence is important not only for LCA, but also for the test in general. That is, examinees who were assigned to Class 4 were causing construct-irrelevant variance in test scores and needed to be identified regardless of which classification method was used.

**Which Approach Should be Adopted for Setting Standards on the ERIT-XA?**

The results suggest examinees with low motivation should be eliminated from the data set through proper screening methods (i.e., motivation filtering; Sundre & Wise, 2003), independent of the standard approach that is selected. If Class 4 is hypothetically excluded from the validity evidence presented in Table 12, there is not a clear-cut choice
about which standard setting method to use. Instead, logistical and practical considerations will be used to advocate for one approach over the other.

The modified Angoff may be viewed as the most preferable approach to standard setting for the ERIT-XA for a few reasons. Chief among them was the intentional process to form groups that meaningfully differed in their performance on the ERIT-XA. Performance category descriptions were created for the modified Angoff to elucidate the cognitive differences believed to exist between groups at the institution where the standard setting was conducted. SMEs were involved in the creation of the performance category descriptions and development of the performance standards as well, which may help stakeholders feel more comfortable with the results than a purely data-driven approach. Also, SMEs who participated in the modified Angoff were content with the final performance standards established after the second round of ratings. Even when presented with impact data that showed the consequences of the performance standards—such as the percentage of students classified into each performance group—SMEs did not desire to make any alterations. It seems reasonable, then, that SMEs were comfortable with the cut scores used to separate performance groups and that they “bought in” to the standards resulting from the modified Angoff. SMEs included faculty, staff, and administration—the major players who will receive ERIT-XA score reports—so it is important they feel comfortable with the process. SMEs may have been less likely to trust the formation of groups if LCA were used instead because it is an abstract process and difficult for a nontechnical audience to understand.

It might seem pretty clear by now that the modified Angoff is the recommended approach to be adopted for the ERIT-XA standard setting. Even though the modified
Angoff is advocated for over the 4-class ordinal LCA, additional work needs to be done. The purpose of this study was not to examine all sources of validity evidence for the standard setting methods. A comprehensive validity analysis of the performance standards and classifications should be investigated.

It is also important to keep in mind the ERIT-XA is only one in a battery of assessments used by MC to assess students’ ethical reasoning skills and abilities. Performance on the ERIT-XA may be a good reflection of a student’s abilities to perform skills foundational to becoming a good ethical reasoner. Although the ERIT-XA represents specific components deemed to be crucial to becoming a complete ethical reasoner, the test does not represent a student’s overall ability to act as a good ethical reasoner. Thus, it is important to connect the instrument with other tools used to assess more advanced ethical reasoning skills. Using a constellation of instruments will facilitate a more complete definition of ethical reasoning and can also be used to create a profile of ethical reasoning for each student. In turn, the nomological network—or the full spectrum of the construct of interest and interrelated aspects—can be captured to enhance valid and meaningful measurement of ethical reasoning.

In addition to collecting additional validity information for the standards and groups, the ERIT-XA development team may wish to further consider the relation between SAT Reading scores and performance on the ERIT-XA. The results suggested the ERIT-XA might be tapping into reading comprehension. If a positive relation between ERIT-XA performance and SAT Reading scores is undesirable, developers of the ERIT-XA may wish to revisit the items of the test and consider simplifying the vocabulary, sentence structure, or other test features related to reading comprehension.
However, if a positive relation between ERIT-XA performance and SAT Reading scores is expected, then such results can be viewed as favorable validity evidence and test alterations may not be necessary.

**Considerations with the Use of Data-driven Approaches**

For the ERIT-XA standard setting, the modified Angoff was advocated because validity evidence was not substantially different between the traditional and data-driven approaches and due to practical reasons. However, some researchers may still desire to use the ordinal LCA to classify examinees into groups. If a data-driven approach is adopted, there are a number of factors that should be considered.

**Consideration 1: Subjectivity is not eliminated.** The first consideration pertains to the subjectivity associated with choosing among solutions from the data-driven models. Consider the circumstances of the present study, in which the primary model comparison within each of the data-driven standard setting methods (i.e., general LCA and ordinal LCA) was the 3-class versus 4-class solution. Although the 4-class solution was ultimately selected as the best-fitting model for both data-driven methods, an argument could be made in favor of the 3-class solution. Class 4 was tiny (about 3% of the sample) in the championed general LCA and ordinal LCA solutions and those students performed poorly on the ERIT-XA. Again, it seems reasonable to question whether the classification of students into Class 4 represents an actual level of achievement/ability on the ERIT-XA, or whether it is suggestive of performance of some other type. The level of subjectivity that may be required by the researcher when choosing a data-driven model to champion is apparent. The decision to champion the 4-class solution for the general and ordinal LCA models was driven by the model-fit
indices in this study. However, a great deal of subjective judgment and interpretation may be required on the part of the researcher when other factors—such as profile plots, meaningfulness of groups, or conflicting model-fit indices—are included in the decision-making process of which solution to champion.

**Consideration 2: Be mindful of the characteristics and size of your sample.** It is also imperative the data sample used for analyses contains participants from the entire spectrum of ability levels on the test. If potential ability levels are excluded from the range of possible cut scores, it is unlikely meaningful and valid interpretations of performance standards and groups will be made. Another important consideration in the application of any statistical technique is whether the size of the sample is large enough to yield trustworthy results. If a sample is considered beyond large enough to yield trustworthy results, it might be divided into subsamples to explore the replicability of the results. For instance, a common technique to evaluate the replicability of a championed solution in LCA is to randomly split the sample in half and conduct analyses on both subsamples (Dziak, Lanza, & Tan, 2014; Lubke, 2010; Masyn, 2013). A relevant question for the current study is whether a sample size of 3,558 students (or 1,779 students, for replicability purposes) is large enough to yield trustworthy results for the LCA models used in the data-driven approaches to standard setting. Unfortunately, it is challenging to determine the appropriate sample size needed to make valid inferences from the results of LCA models (Dziak et al., 2014; Gudicha, Tekle, & Vermunt, 2016; Masyn, 2013). In contrast with other statistical models (e.g., ANOVA, linear regression), it is difficult to label a specific sample size that suits all studies because power analysis in LCA depends on several population and study design characteristics (Muthén & Muthén,
2002; Tekle, Gudicha, & Vermunt, 2016). First, power in LCA means something different than the traditional understanding of statistical power. A power analysis in LCA is used to explore the sample size needed to find the correct number of classes that exist in the population. Use of an inadequate sample size in LCA diminishes the ability to detect classes that may be important but are not highly prevalent (Berlin, Williams, & Parra, 2014; Dziak et al., 2014; Masyn, 2013). In fact, Tekle et al. (2016) concluded it is erroneous to declare a specific sample size (e.g., 200 or 500 examinees) as indicative of sufficient power for all LCA studies.

Other factors that must be considered for power analysis in LCA include class weights, the number of classes, the number of observed indicator variables (e.g., items on the ERIT-XA), and separation level between classes (Tekle et al., 2016). Based on results from a simulation study, Wurpts and Geiser (2014) found a higher number of indicators and higher quality indicators can compensate for small sample size in LCA. Gudicha et al. (2016) also performed simulation studies to examine the power and sample size computations for latent class models. Their results suggested a smaller number of latent classes, larger number of indicator variables, stronger associations between classes and indicators, more equal class sizes, and greater separation between different latent classes was also shown to improve power in latent class models (Gudicha et al., 2016). A major factor cited by Tekle et al. (2016) as having influence on statistical power in LCA is the extent to which classes are separated. When high separation is found between classes, sample size and the number of indicators can be reduced, compared to when class separation is low (Tekle et al., 2016). Ultimately, an array of considerations must be
taken into account when evaluating the sample size necessary to obtain sufficient power for LCA.

**Consideration 3: Ordered classes are not guaranteed.** The finding of ordered classes in the current study—while certainly beneficial and aligned with the intention to yield groups that differ quantitatively—will not always occur with the use of a general LCA. The ERIT-XA had undergone extensive test development prior to the data-driven standard setting using the general LCA models. If the factor structure of the test had not been properly examined, unordered classes may have emerged from the general LCA. If unordered groups are encountered after estimating a general LCA model, there are typically two explanations. Unordered groups suggest: (1) standards may be inappropriate for a test, or (2) a test is multidimensional and standards need to be set for each dimension or subscale of the test. In either event, the researcher will likely have to revisit the test development process before establishing appropriate performance standards.

**Consideration 4: There are different ways to assign examinees to classes.** Finally, a decision about how to classify examinees into latent classes also needs to be made. Because examinees are assigned to classes in LCA, different groups can be formed without establishing specific cut scores. As previously mentioned, examinees can be classified into groups based on their largest posterior probability of class membership (i.e., using modal assignment). However, the creation of groups without knowledge of the scores that were used to separate performance levels may make stakeholders leery of the results. If cut scores are desired, the overlap between adjacent test score distributions may be used. For instance, the median of two overlapping groups can be calculated to
establish a specific cut score. Another option may be to use the minimum score of a higher group and maximum score of a lower group to construct a range of values that serve as cut scores.

**Future Research**

As argued in earlier chapters, perhaps the best approach to standard setting is one that integrates the judgments of SMEs and data-driven classification techniques. The Angoff LCA is a promising method that combines aspects of traditional and data-driven approaches, but it might be too restrictive in some situations for proper estimation of the model parameters. Rather than using SMEs’ ratings to constrain item difficulties to fall within particular boundaries, other prior information may make convergence more feasible. For instance, asking SMEs to provide the average probability of correct response for each performance category by item would allow researchers to specify a prior distribution for the item difficulties using a Bayesian LCA framework. Cut scores for such a method would be difficult to calculate, however, because SMEs’ ratings would be based on expected performance of students within each category rather than that of borderline examinees. Additionally, Bayesian software and know-how are also required to estimate such a model. Although there may be challenges associated with a Bayesian Angoff LCA, its consideration is worthy of attention.

Future research should also investigate how the various components of a power analysis in LCA affect convergence for integrated models. A large number of indicators were included in the Angoff LCA (50 items), but a large number of constraints also needed to be applied to the items of the championed 4-class ordinal LCA solution. Additional studies examining the factors that affect convergence of integrated LCA
models are needed. There still remains great scrutiny about the traditional and data-driven methods to standard setting, and the discovery of alternative ways to set standards can certainly provide a benefit.

Continued research on the performance standards for the ERIT-XA and related ethical reasoning assessment should be considered as well. The major premise of the ERIT-XA standard setting was to gather more nuanced detail regarding students’ ethical reasoning development. Identifying the correct Key Question to an ethical dilemma can be considered a foundational skill for students developing in ethical reasoning. Setting performance standards on the ERIT-XA was not intended to determine whether students are competent ethical reasoners; rather, the intent was to identify mastery of foundational skills of ethical reasoning. Performance groups on the ERIT-XA that resulted from the modified Angoff can certainly fulfill that purpose. An even greater perspective of ethical reasoning at JMU can be facilitated by creating student profiles based on performance on related ethical reasoning assessments administered through the MC. An idea may be to reward students for achieving performance aligned with Proficient and Advanced ethical reasoning profiles, in general, and may also serve as a model for other institutions.

**Conclusions**

Subjectivity is a common problem when setting performance standards using traditional (Cizek, 2012; Hambleton, 1978; Kane, 2001a; Popham, 1978) or data-driven methods (Sireci et al., 1999). If the results from a standard setting are deemed too subjective to provide valuable meaning, faculty and other stakeholders may exhibit a lack of trust in the performance labels assigned to a test. Multiple sources of information (SMEs’ judgment and data) were combined from the traditional and statistical standard
setting methods in this dissertation to form an “integrated” approach. The rationale behind creating the Angoff LCA was to produce improved ratings that enable researchers to make more informed classification decisions. Although the Angoff LCA failed to converge, further research should be conducted before dismissing it and other integrated models as an ineffective means to set performance standards on a test.

Of the data-driven models that did converge, selection of a championed class solution required subjective interpretation within and across methods due to conflicting evidence, such as inconsistent model-fit indicators. It was also tricky determining which criteria were most suitable to compare the results from the different standard setting approaches. Traditional standard setting methods will always have the advantage of heavily involving SMEs in multiple facets of the process. Data-driven approaches, on the other hand, suggest what model fits the data best. In this study, LCA was used to indicate the number of nature of groups that were best characterized by the data. However, other data-driven classification methods, such as cluster analysis or mixture Rasch models, may have produced different interpretations and conclusions. Further exploration and guidance about how to properly compare different standard setting approaches should be explored.

Ultimately, the traditional approach was espoused as the most appropriate standard setting technique for the ERIT-XA, but it only represents the results of one test. There are many different ways to classify students into groups using traditional, data-driven, or integrated approaches. Perhaps the Angoff LCA works better as a standard setting device used for other tests, or the ordinal LCA produces more favorable classes
and validity evidence in a different design. Further studies should be conducted to examine the usefulness of data-driven and integrated approaches.
Table 1

*Latent Class Model Comparisons*

<table>
<thead>
<tr>
<th></th>
<th>Free parameters</th>
<th>LL</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-class</td>
<td>12</td>
<td>-1184.81</td>
<td>2393.62</td>
<td>2432.84</td>
</tr>
<tr>
<td>2-class</td>
<td>25</td>
<td>-1125.60</td>
<td>2301.19</td>
<td>2382.89</td>
</tr>
<tr>
<td>3-class</td>
<td>36</td>
<td>-1117.66</td>
<td>2307.25</td>
<td>2424.90</td>
</tr>
</tbody>
</table>

*Note:* LL = Log likelihood; AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria. Table adapted from “Using latent class analysis to set academic performance standards” by R. S. Brown, 2007, *Educational Assessment, 12,* 283-301.
Table 2

*Brief Description of 8KQ Provided to Students who take the ERIT-XA*

<table>
<thead>
<tr>
<th>Key Question</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Empathy</td>
<td>How would someone respond if s/he cared deeply about those involved?</td>
</tr>
<tr>
<td>(2) Fairness</td>
<td>How can someone act equitably and balance all interests?</td>
</tr>
<tr>
<td>(3) Character</td>
<td>What actions will help someone become his/her ideal self?</td>
</tr>
<tr>
<td>(4) Liberty</td>
<td>What principles of freedom and personal autonomy apply?</td>
</tr>
<tr>
<td>(5) Rights</td>
<td>What rights (e.g., innate, legal, social) apply?</td>
</tr>
<tr>
<td>(6) Responsibilities</td>
<td>What duties and obligations apply?</td>
</tr>
<tr>
<td>(7) Outcomes</td>
<td>What are the short-term and long-term outcomes of possible actions?</td>
</tr>
<tr>
<td>(8) Authority</td>
<td>What do legitimate authorities (e.g., experts, the law, one’s god[s]) expect of someone?</td>
</tr>
<tr>
<td>Table 3</td>
<td>ERIT-XA Performance Category Descriptions</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td><strong>Developing</strong></td>
<td><strong>Proficient</strong></td>
</tr>
<tr>
<td>Students who are <em>Developing</em> in their ability to identify the Key Question (KQ) most consistent with a decision/rationale:</td>
<td>Students who are <em>Proficient</em> in their ability to identify the Key Question (KQ) most consistent with a decision/rationale:</td>
</tr>
</tbody>
</table>

**A. Knowledge and Understanding**

- Demonstrate an uneven understanding of the 8 KQs.
- Exhibit a fairly strong grasp on most KQs that are easily understandable given the short descriptions provided on the test (e.g., Empathy, Fairness, Character).
- Often struggle with KQs that require more than just the description provided to understand (e.g., Liberty, Rights, Responsibilities).

- Demonstrate a firm understanding of the 8 KQs.
- Exhibit a strong grasp on all KQs that are easily understandable given the short descriptions provided on the test (e.g., Empathy, Fairness, Character).
- Sometimes struggle with KQs that require more than just the description provided to understand (e.g., Liberty, Rights, Responsibilities).

- Demonstrate a complex, nuanced, and sophisticated understanding of the 8 KQs.
- Exhibit a strong grasp on all KQs, regardless of the description provided on the test.
- Do not struggle with KQs that require more than just the description provided to understand (e.g., Liberty, Rights, Responsibilities).

**B. Identification**

- Can identify the KQ most consistent with a decision/rationale when there is one obvious choice available in the scenario and the most consistent KQ is one that is easily understandable.
- Often struggle to identify the most consistent KQ when there are multiple possibilities, particularly when the possibilities include KQs that require more than just the description provided to understand.

- Can identify the KQ most consistent with a decision/rationale when there is one obvious choice available in the scenario and the most consistent KQ is one that is easily understandable.
- Sometimes struggle to identify the most consistent KQ when there are multiple possibilities, particularly when the possibilities include KQs that require more than just the description provided to understand.

- Can identify the KQ most consistent with a decision/rationale when there is one obvious choice available in the scenario.
- Can also identify the KQ most consistent with a decision/rationale when there are multiple possibilities, regardless of whether the possibilities include KQs that require more than just the description provided to understand.
|   |   | • Are often able to discriminate amongst multiple KQs to identify the one most consistent with a decision/rationale. |
Table 4  
*Student Status, Testing Context, and Sample Size for ERIT-XA data*

<table>
<thead>
<tr>
<th>Student Status</th>
<th>Context &amp; Date</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incoming Freshmen</td>
<td>Assessment Day, Fall 2013</td>
<td>401</td>
<td>11.27</td>
</tr>
<tr>
<td>Incoming Freshmen</td>
<td>Assessment Day, Fall 2014</td>
<td>211</td>
<td>5.93</td>
</tr>
<tr>
<td>Incoming Freshmen</td>
<td>Assessment Day, Fall 2015</td>
<td>465</td>
<td>13.07</td>
</tr>
<tr>
<td>Incoming Freshmen</td>
<td>Assessment Day, Fall 2016</td>
<td>403</td>
<td>11.33</td>
</tr>
<tr>
<td>Variable</td>
<td>Madison Collaborative, Fall 2016</td>
<td>242</td>
<td>6.80</td>
</tr>
<tr>
<td>Sophomore/Junior</td>
<td>Assessment Day, Spring 2014</td>
<td>789</td>
<td>22.18</td>
</tr>
<tr>
<td>Sophomore/Junior</td>
<td>Assessment Day, Spring 2015</td>
<td>639</td>
<td>17.96</td>
</tr>
<tr>
<td>Sophomore/Junior</td>
<td>Assessment Day, Spring 2016</td>
<td>408</td>
<td>11.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3558</td>
<td>100.00</td>
</tr>
</tbody>
</table>
## Table 5

**ERIT-XA Performance Standards by Round and Rater**

<table>
<thead>
<tr>
<th>Round</th>
<th>Rater</th>
<th>MP Raw Score Standard</th>
<th>MA Raw Score Standard</th>
<th>MP Percent-Correct Standard</th>
<th>MA Percent-Correct Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>23.89</td>
<td>34.58</td>
<td>47.78%</td>
<td>69.16%</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>24.81</td>
<td>42.51</td>
<td>49.62%</td>
<td>85.02%</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>26.30</td>
<td>36.29</td>
<td>52.60%</td>
<td>72.58%</td>
</tr>
<tr>
<td>4</td>
<td>29.95</td>
<td>40.35</td>
<td>59.90%</td>
<td>80.70%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30.35</td>
<td>41.80</td>
<td>60.70%</td>
<td>83.60%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>31.56</td>
<td>42.59</td>
<td>63.12%</td>
<td>85.18%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>32.18</td>
<td>36.57</td>
<td>64.36%</td>
<td>73.14%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>35.93</td>
<td>44.94</td>
<td>71.86%</td>
<td>89.88%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>37.12</td>
<td>46.79</td>
<td>74.24%</td>
<td>93.58%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>37.90</td>
<td>44.17</td>
<td>75.80%</td>
<td>88.34%</td>
<td></td>
</tr>
</tbody>
</table>

|            |       |                       |                       |                            |                            |
| Mdn        | 30.96 | 42.16                 | 61.91%               | 84.31%                      |                             |
| SD         | 4.98  | 4.05                  | 9.97%                | 8.11%                       |                             |
| Min        | 23.89 | 34.58                 | 47.78%               | 69.16%                      |                             |
| Max        | 37.90 | 46.79                 | 75.80%               | 93.58%                      |                             |

<table>
<thead>
<tr>
<th>Round</th>
<th>Rater</th>
<th>MP Raw Score Standard</th>
<th>MA Raw Score Standard</th>
<th>MP Percent-Correct Standard</th>
<th>MA Percent-Correct Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>26.25</td>
<td>42.29</td>
<td>52.50%</td>
<td>84.58%</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>26.55</td>
<td>36.71</td>
<td>53.10%</td>
<td>73.42%</td>
</tr>
<tr>
<td>3</td>
<td>26.77</td>
<td>36.56</td>
<td>53.54%</td>
<td>73.12%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>27.95</td>
<td>40.65</td>
<td>55.90%</td>
<td>81.30%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30.25</td>
<td>40.66</td>
<td>60.50%</td>
<td>81.32%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>30.70</td>
<td>39.55</td>
<td>61.40%</td>
<td>79.10%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>31.00</td>
<td>37.75</td>
<td>62.00%</td>
<td>75.50%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>32.30</td>
<td>43.30</td>
<td>64.60%</td>
<td>86.60%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>34.74</td>
<td>44.96</td>
<td>69.48%</td>
<td>89.92%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>35.87</td>
<td>45.37</td>
<td>71.74%</td>
<td>90.74%</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Mdn</td>
<td>30.48</td>
<td>40.66</td>
<td></td>
<td>60.95%</td>
<td>81.31%</td>
</tr>
<tr>
<td>SD</td>
<td>3.40</td>
<td>3.20</td>
<td></td>
<td>6.79%</td>
<td>6.40%</td>
</tr>
<tr>
<td>Min</td>
<td>26.25</td>
<td>36.56</td>
<td></td>
<td>52.50%</td>
<td>73.12%</td>
</tr>
<tr>
<td>Max</td>
<td>35.87</td>
<td>45.37</td>
<td></td>
<td>71.74%</td>
<td>90.74%</td>
</tr>
</tbody>
</table>

*Note. MP = Minimally Proficient, MA = Minimally Advanced*
Table 6
Median Ratings by Borderline Group for ERIT-XA Items (Round 1)

<table>
<thead>
<tr>
<th>Item</th>
<th>Minimally Proficient</th>
<th>SD</th>
<th>Minimally Advanced</th>
<th>SD</th>
<th>Difference Between MA/MP Ratings</th>
<th>Item Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70.50</td>
<td>8.00</td>
<td>91.00</td>
<td>13.73</td>
<td>20.50</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>67.50</td>
<td>20.20</td>
<td>82.50</td>
<td>21.78</td>
<td>15.00</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>68.00</td>
<td>21.66</td>
<td>83.00</td>
<td>19.58</td>
<td>15.00</td>
<td>0.72</td>
</tr>
<tr>
<td>4</td>
<td>60.50</td>
<td>22.02</td>
<td>80.50</td>
<td>14.90</td>
<td>20.00</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>70.00</td>
<td>16.28</td>
<td>87.50</td>
<td>8.83</td>
<td>17.50</td>
<td>0.57</td>
</tr>
<tr>
<td>6</td>
<td>70.00</td>
<td>8.11</td>
<td>90.00</td>
<td>6.09</td>
<td>20.00</td>
<td>0.54</td>
</tr>
<tr>
<td>7</td>
<td>50.00</td>
<td>8.85</td>
<td>75.00</td>
<td>10.60</td>
<td>25.00</td>
<td>0.54</td>
</tr>
<tr>
<td>8</td>
<td>50.00</td>
<td>10.83</td>
<td>80.00</td>
<td>9.62</td>
<td>30.00</td>
<td>0.58</td>
</tr>
<tr>
<td>9</td>
<td>80.00</td>
<td>6.60</td>
<td>95.00</td>
<td>3.72</td>
<td>15.00</td>
<td>0.92</td>
</tr>
<tr>
<td>10</td>
<td>52.50</td>
<td>17.99</td>
<td>80.00</td>
<td>14.53</td>
<td>27.50</td>
<td>0.47</td>
</tr>
<tr>
<td>11</td>
<td>52.50</td>
<td>20.08</td>
<td>75.00</td>
<td>19.37</td>
<td>22.50</td>
<td>0.65</td>
</tr>
<tr>
<td>12</td>
<td>80.50</td>
<td>7.18</td>
<td>95.50</td>
<td>4.38</td>
<td>15.00</td>
<td>0.94</td>
</tr>
<tr>
<td>13</td>
<td>67.50</td>
<td>12.53</td>
<td>87.50</td>
<td>6.96</td>
<td>20.00</td>
<td>0.60</td>
</tr>
<tr>
<td>14</td>
<td>51.00</td>
<td>17.33</td>
<td>75.00</td>
<td>12.12</td>
<td>24.00</td>
<td>0.50</td>
</tr>
<tr>
<td>15</td>
<td>75.00</td>
<td>16.10</td>
<td>92.50</td>
<td>9.67</td>
<td>17.50</td>
<td>0.74</td>
</tr>
<tr>
<td>16</td>
<td>60.00</td>
<td>14.77</td>
<td>83.50</td>
<td>9.76</td>
<td>23.50</td>
<td>0.73</td>
</tr>
<tr>
<td>17</td>
<td>72.50</td>
<td>15.67</td>
<td>90.00</td>
<td>10.60</td>
<td>17.50</td>
<td>0.88</td>
</tr>
<tr>
<td>18</td>
<td>55.00</td>
<td>22.84</td>
<td>77.50</td>
<td>24.78</td>
<td>22.50</td>
<td>0.65</td>
</tr>
<tr>
<td>19</td>
<td>65.00</td>
<td>15.54</td>
<td>87.50</td>
<td>11.32</td>
<td>22.50</td>
<td>0.72</td>
</tr>
<tr>
<td>20</td>
<td>67.50</td>
<td>17.09</td>
<td>88.50</td>
<td>12.63</td>
<td>21.00</td>
<td>0.63</td>
</tr>
<tr>
<td>21</td>
<td>80.00</td>
<td>6.74</td>
<td>95.00</td>
<td>4.99</td>
<td>15.00</td>
<td>0.95</td>
</tr>
<tr>
<td>22</td>
<td>57.50</td>
<td>21.63</td>
<td>82.50</td>
<td>22.20</td>
<td>25.00</td>
<td>0.70</td>
</tr>
<tr>
<td>23</td>
<td>67.50</td>
<td>17.70</td>
<td>86.00</td>
<td>16.84</td>
<td>18.50</td>
<td>0.77</td>
</tr>
<tr>
<td>24</td>
<td>52.50</td>
<td>18.44</td>
<td>80.50</td>
<td>16.12</td>
<td>28.00</td>
<td>0.53</td>
</tr>
<tr>
<td>25</td>
<td>70.00</td>
<td>9.44</td>
<td>94.00</td>
<td>7.09</td>
<td>24.00</td>
<td>0.83</td>
</tr>
<tr>
<td>26</td>
<td>78.00</td>
<td>11.25</td>
<td>95.00</td>
<td>4.37</td>
<td>17.00</td>
<td>0.50</td>
</tr>
<tr>
<td>27</td>
<td>80.00</td>
<td>8.23</td>
<td>95.00</td>
<td>6.81</td>
<td>15.00</td>
<td>0.89</td>
</tr>
<tr>
<td>28</td>
<td>67.50</td>
<td>15.35</td>
<td>87.50</td>
<td>13.23</td>
<td>20.00</td>
<td>0.91</td>
</tr>
<tr>
<td>29</td>
<td>50.00</td>
<td>16.69</td>
<td>77.50</td>
<td>17.03</td>
<td>27.50</td>
<td>0.57</td>
</tr>
<tr>
<td>30</td>
<td>42.50</td>
<td>19.62</td>
<td>67.50</td>
<td>17.61</td>
<td>25.00</td>
<td>0.50</td>
</tr>
<tr>
<td>31</td>
<td>72.50</td>
<td>11.61</td>
<td>93.50</td>
<td>6.18</td>
<td>21.00</td>
<td>0.91</td>
</tr>
<tr>
<td>32</td>
<td>70.00</td>
<td>12.44</td>
<td>90.50</td>
<td>5.67</td>
<td>20.50</td>
<td>0.61</td>
</tr>
<tr>
<td>33</td>
<td>65.00</td>
<td>14.89</td>
<td>90.00</td>
<td>8.93</td>
<td>25.00</td>
<td>0.58</td>
</tr>
<tr>
<td>34</td>
<td>76.00</td>
<td>9.52</td>
<td>95.00</td>
<td>7.53</td>
<td>19.00</td>
<td>0.85</td>
</tr>
<tr>
<td>35</td>
<td>79.00</td>
<td>7.69</td>
<td>95.00</td>
<td>4.53</td>
<td>16.00</td>
<td>0.92</td>
</tr>
<tr>
<td>36</td>
<td>50.00</td>
<td>19.02</td>
<td>72.50</td>
<td>17.86</td>
<td>22.50</td>
<td>0.53</td>
</tr>
<tr>
<td>37</td>
<td>40.00</td>
<td>17.85</td>
<td>71.50</td>
<td>18.04</td>
<td>31.50</td>
<td>0.87</td>
</tr>
<tr>
<td>38</td>
<td>70.00</td>
<td>22.44</td>
<td>88.00</td>
<td>20.74</td>
<td>18.00</td>
<td>0.86</td>
</tr>
<tr>
<td>39</td>
<td>42.50</td>
<td>21.37</td>
<td>74.00</td>
<td>22.01</td>
<td>31.50</td>
<td>0.17</td>
</tr>
<tr>
<td>40</td>
<td>66.50</td>
<td>21.73</td>
<td>82.50</td>
<td>16.65</td>
<td>16.00</td>
<td>0.78</td>
</tr>
<tr>
<td>41</td>
<td>75.00</td>
<td>13.23</td>
<td>90.00</td>
<td>7.65</td>
<td>15.00</td>
<td>0.61</td>
</tr>
<tr>
<td>42</td>
<td>58.00</td>
<td>16.23</td>
<td>80.00</td>
<td>20.19</td>
<td>22.00</td>
<td>0.37</td>
</tr>
<tr>
<td>43</td>
<td>67.50</td>
<td>20.88</td>
<td>87.50</td>
<td>17.60</td>
<td>20.00</td>
<td>0.73</td>
</tr>
<tr>
<td>44</td>
<td>65.00</td>
<td>17.89</td>
<td>82.50</td>
<td>11.35</td>
<td>17.50</td>
<td>0.60</td>
</tr>
<tr>
<td>45</td>
<td>74.50</td>
<td>14.54</td>
<td>91.50</td>
<td>6.67</td>
<td>17.00</td>
<td>0.74</td>
</tr>
<tr>
<td>46</td>
<td>51.50</td>
<td>21.62</td>
<td>77.50</td>
<td>26.43</td>
<td>26.00</td>
<td>0.22</td>
</tr>
<tr>
<td>47</td>
<td>63.00</td>
<td>16.27</td>
<td>82.50</td>
<td>16.51</td>
<td>19.50</td>
<td>0.89</td>
</tr>
<tr>
<td>48</td>
<td>70.00</td>
<td>19.76</td>
<td>85.00</td>
<td>14.35</td>
<td>15.00</td>
<td>0.94</td>
</tr>
<tr>
<td>49</td>
<td>72.50</td>
<td>11.54</td>
<td>91.00</td>
<td>10.39</td>
<td>18.50</td>
<td>0.90</td>
</tr>
<tr>
<td>50</td>
<td>52.50</td>
<td>22.54</td>
<td>80.00</td>
<td>18.87</td>
<td>27.50</td>
<td>0.70</td>
</tr>
<tr>
<td>Item</td>
<td>Minimally Proficient</td>
<td>SD</td>
<td>Minimally Advanced</td>
<td>SD</td>
<td>Difference Between MA/MP Ratings</td>
<td>Item Difficulty</td>
</tr>
<tr>
<td>------</td>
<td>----------------------</td>
<td>------</td>
<td>--------------------</td>
<td>------</td>
<td>----------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>1</td>
<td>68.50</td>
<td>10.52</td>
<td>85.00</td>
<td>7.20</td>
<td>16.50</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>57.50</td>
<td>15.30</td>
<td>80.00</td>
<td>12.61</td>
<td>22.50</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>62.00</td>
<td>11.66</td>
<td>81.00</td>
<td>12.50</td>
<td>19.00</td>
<td>0.72</td>
</tr>
<tr>
<td>4</td>
<td>65.00</td>
<td>15.83</td>
<td>80.50</td>
<td>9.67</td>
<td>15.50</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>70.50</td>
<td>9.10</td>
<td>90.00</td>
<td>5.10</td>
<td>19.50</td>
<td>0.57</td>
</tr>
<tr>
<td>6</td>
<td>67.50</td>
<td>8.90</td>
<td>87.50</td>
<td>6.18</td>
<td>20.00</td>
<td>0.54</td>
</tr>
<tr>
<td>7</td>
<td>50.00</td>
<td>9.35</td>
<td>80.00</td>
<td>8.68</td>
<td>30.00</td>
<td>0.54</td>
</tr>
<tr>
<td>8</td>
<td>55.00</td>
<td>10.53</td>
<td>80.00</td>
<td>10.52</td>
<td>25.00</td>
<td>0.58</td>
</tr>
<tr>
<td>9</td>
<td>74.50</td>
<td>7.78</td>
<td>90.00</td>
<td>6.95</td>
<td>15.50</td>
<td>0.92</td>
</tr>
<tr>
<td>10</td>
<td>52.50</td>
<td>16.24</td>
<td>77.50</td>
<td>12.70</td>
<td>25.00</td>
<td>0.47</td>
</tr>
<tr>
<td>11</td>
<td>57.50</td>
<td>17.44</td>
<td>77.50</td>
<td>14.23</td>
<td>20.00</td>
<td>0.65</td>
</tr>
<tr>
<td>12</td>
<td>80.00</td>
<td>5.96</td>
<td>95.00</td>
<td>5.42</td>
<td>15.00</td>
<td>0.94</td>
</tr>
<tr>
<td>13</td>
<td>66.00</td>
<td>12.81</td>
<td>89.50</td>
<td>7.72</td>
<td>23.50</td>
<td>0.60</td>
</tr>
<tr>
<td>14</td>
<td>50.00</td>
<td>11.80</td>
<td>75.00</td>
<td>10.78</td>
<td>25.00</td>
<td>0.50</td>
</tr>
<tr>
<td>15</td>
<td>75.00</td>
<td>13.66</td>
<td>90.00</td>
<td>7.18</td>
<td>15.00</td>
<td>0.74</td>
</tr>
<tr>
<td>16</td>
<td>60.00</td>
<td>12.46</td>
<td>82.50</td>
<td>9.08</td>
<td>22.50</td>
<td>0.73</td>
</tr>
<tr>
<td>17</td>
<td>70.00</td>
<td>9.48</td>
<td>90.00</td>
<td>7.35</td>
<td>20.00</td>
<td>0.88</td>
</tr>
<tr>
<td>18</td>
<td>55.00</td>
<td>19.70</td>
<td>82.50</td>
<td>18.57</td>
<td>27.50</td>
<td>0.65</td>
</tr>
<tr>
<td>19</td>
<td>62.50</td>
<td>11.60</td>
<td>87.50</td>
<td>8.03</td>
<td>25.00</td>
<td>0.72</td>
</tr>
<tr>
<td>20</td>
<td>60.50</td>
<td>14.96</td>
<td>88.00</td>
<td>11.67</td>
<td>27.50</td>
<td>0.63</td>
</tr>
<tr>
<td>21</td>
<td>77.50</td>
<td>6.37</td>
<td>95.00</td>
<td>6.62</td>
<td>17.50</td>
<td>0.95</td>
</tr>
<tr>
<td>22</td>
<td>50.00</td>
<td>13.84</td>
<td>77.50</td>
<td>12.19</td>
<td>27.50</td>
<td>0.70</td>
</tr>
<tr>
<td>23</td>
<td>67.50</td>
<td>18.81</td>
<td>89.50</td>
<td>15.17</td>
<td>22.00</td>
<td>0.77</td>
</tr>
<tr>
<td>24</td>
<td>47.50</td>
<td>10.96</td>
<td>79.00</td>
<td>13.45</td>
<td>31.50</td>
<td>0.53</td>
</tr>
<tr>
<td>25</td>
<td>70.50</td>
<td>3.71</td>
<td>91.50</td>
<td>4.19</td>
<td>21.00</td>
<td>0.83</td>
</tr>
<tr>
<td>26</td>
<td>74.00</td>
<td>8.63</td>
<td>92.00</td>
<td>5.12</td>
<td>18.00</td>
<td>0.50</td>
</tr>
<tr>
<td>27</td>
<td>75.00</td>
<td>7.11</td>
<td>94.50</td>
<td>7.01</td>
<td>19.50</td>
<td>0.89</td>
</tr>
<tr>
<td>28</td>
<td>65.00</td>
<td>12.51</td>
<td>85.50</td>
<td>10.11</td>
<td>20.50</td>
<td>0.91</td>
</tr>
<tr>
<td>29</td>
<td>45.00</td>
<td>18.27</td>
<td>72.50</td>
<td>21.81</td>
<td>27.50</td>
<td>0.57</td>
</tr>
<tr>
<td>30</td>
<td>43.00</td>
<td>19.45</td>
<td>70.00</td>
<td>16.73</td>
<td>27.00</td>
<td>0.50</td>
</tr>
<tr>
<td>31</td>
<td>67.50</td>
<td>10.12</td>
<td>90.00</td>
<td>4.95</td>
<td>22.50</td>
<td>0.91</td>
</tr>
<tr>
<td>32</td>
<td>68.00</td>
<td>6.74</td>
<td>90.00</td>
<td>3.98</td>
<td>22.00</td>
<td>0.61</td>
</tr>
<tr>
<td>33</td>
<td>62.50</td>
<td>8.53</td>
<td>86.50</td>
<td>7.37</td>
<td>24.00</td>
<td>0.58</td>
</tr>
<tr>
<td>34</td>
<td>73.50</td>
<td>9.80</td>
<td>90.00</td>
<td>8.66</td>
<td>16.50</td>
<td>0.85</td>
</tr>
<tr>
<td>35</td>
<td>70.50</td>
<td>7.11</td>
<td>90.00</td>
<td>4.14</td>
<td>19.50</td>
<td>0.92</td>
</tr>
<tr>
<td>36</td>
<td>42.50</td>
<td>16.19</td>
<td>70.00</td>
<td>15.15</td>
<td>27.50</td>
<td>0.53</td>
</tr>
<tr>
<td>37</td>
<td>60.50</td>
<td>16.43</td>
<td>78.50</td>
<td>17.00</td>
<td>18.00</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>38</td>
<td>65.00</td>
<td>18.09</td>
<td>85.00</td>
<td>17.55</td>
<td>20.00</td>
<td>0.86</td>
</tr>
<tr>
<td>39</td>
<td>32.50</td>
<td>15.89</td>
<td>67.50</td>
<td>15.69</td>
<td>35.00</td>
<td>0.17</td>
</tr>
<tr>
<td>40</td>
<td>61.00</td>
<td>14.98</td>
<td>82.50</td>
<td>12.21</td>
<td>21.50</td>
<td>0.78</td>
</tr>
<tr>
<td>41</td>
<td>63.00</td>
<td>11.95</td>
<td>85.50</td>
<td>8.98</td>
<td>22.50</td>
<td>0.61</td>
</tr>
<tr>
<td>42</td>
<td>52.50</td>
<td>12.18</td>
<td>81.50</td>
<td>11.15</td>
<td>29.00</td>
<td>0.37</td>
</tr>
<tr>
<td>43</td>
<td>65.00</td>
<td>12.98</td>
<td>85.00</td>
<td>12.91</td>
<td>20.00</td>
<td>0.73</td>
</tr>
<tr>
<td>44</td>
<td>60.00</td>
<td>8.84</td>
<td>82.50</td>
<td>10.35</td>
<td>22.50</td>
<td>0.60</td>
</tr>
<tr>
<td>45</td>
<td>70.00</td>
<td>12.74</td>
<td>90.00</td>
<td>10.60</td>
<td>20.00</td>
<td>0.74</td>
</tr>
<tr>
<td>46</td>
<td>37.50</td>
<td>14.62</td>
<td>58.00</td>
<td>20.95</td>
<td>20.50</td>
<td>0.22</td>
</tr>
<tr>
<td>47</td>
<td>59.50</td>
<td>12.00</td>
<td>80.00</td>
<td>12.81</td>
<td>20.50</td>
<td>0.89</td>
</tr>
<tr>
<td>48</td>
<td>67.50</td>
<td>14.87</td>
<td>82.50</td>
<td>13.69</td>
<td>15.00</td>
<td>0.94</td>
</tr>
<tr>
<td>49</td>
<td>70.50</td>
<td>8.86</td>
<td>89.00</td>
<td>8.35</td>
<td>18.50</td>
<td>0.90</td>
</tr>
<tr>
<td>50</td>
<td>50.00</td>
<td>15.66</td>
<td>73.00</td>
<td>11.27</td>
<td>23.00</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Table 8
Fit Indices and Entropy for General LCA, Ordinal LCA, and Factor Models

<table>
<thead>
<tr>
<th></th>
<th># of paras</th>
<th>LL</th>
<th>BIC</th>
<th>SSABIC</th>
<th>SIC</th>
<th>BF</th>
<th>LMR p</th>
<th>BLRT p</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General LCA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-class</td>
<td>50</td>
<td>-96321.39</td>
<td>193051.62</td>
<td>192892.74</td>
<td>-96525.81</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1.000</td>
</tr>
<tr>
<td>2-class</td>
<td>101</td>
<td>-90976.91</td>
<td>182779.69</td>
<td>182458.76</td>
<td>-91389.84</td>
<td>0.00</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.847</td>
</tr>
<tr>
<td>3-class</td>
<td>152</td>
<td>-89160.50</td>
<td>179563.90</td>
<td>179080.92</td>
<td>-89781.95</td>
<td>0.00</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>0.853</td>
</tr>
<tr>
<td>4-class</td>
<td>203</td>
<td>-88729.63</td>
<td>179119.17</td>
<td>178474.14</td>
<td>-89559.59</td>
<td>0.00</td>
<td>0.0023</td>
<td>&lt;.0001</td>
<td>0.815</td>
</tr>
<tr>
<td>5-class</td>
<td>254</td>
<td>-88525.83</td>
<td>179128.61</td>
<td>178321.53</td>
<td>-89564.31</td>
<td>112.17</td>
<td>0.14</td>
<td>--</td>
<td>0.733</td>
</tr>
<tr>
<td><strong>Ordinal LCA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-class</td>
<td>101</td>
<td>-90976.91</td>
<td>182779.69</td>
<td>182458.76</td>
<td>-91389.84</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>0.847</td>
</tr>
<tr>
<td>3-class</td>
<td>152</td>
<td>-89160.50</td>
<td>179563.90</td>
<td>179080.92</td>
<td>-89781.95</td>
<td>0.00</td>
<td>--</td>
<td>&lt;.0001</td>
<td>0.853</td>
</tr>
<tr>
<td>4-class</td>
<td>203</td>
<td>-88729.72</td>
<td>179119.36</td>
<td>178474.33</td>
<td>-89559.68</td>
<td>0.00</td>
<td>--</td>
<td>&lt;.0001</td>
<td>0.815</td>
</tr>
<tr>
<td><strong>Factor Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-factor</td>
<td>100</td>
<td>-88829.40</td>
<td>178476.49</td>
<td>178158.75</td>
<td>-89238.25</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

*Note.* # of paras = Number of parameters estimated, LL = Log-likelihood, BIC = Bayesian information criterion, SSABIC = Sample size adjusted Bayesian information criterion, BF = Bayes Factor, LMR p = Lo-Mendell-Rubin p-value, BLRT p = Bootstrap likelihood ratio p-value.
Table 9

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advanced</strong></td>
<td>698</td>
<td>43.12</td>
<td>1.85</td>
<td>41</td>
<td>50</td>
</tr>
<tr>
<td><strong>Proficient</strong></td>
<td>1932</td>
<td>35.64</td>
<td>2.80</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td><strong>Developing</strong></td>
<td>928</td>
<td>24.28</td>
<td>6.05</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td><strong>Class 1</strong></td>
<td>1175</td>
<td>41.39</td>
<td>2.63</td>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td><strong>Class 2</strong></td>
<td>1720</td>
<td>33.73</td>
<td>2.96</td>
<td>26</td>
<td>41</td>
</tr>
<tr>
<td><strong>Class 3</strong></td>
<td>540</td>
<td>24.78</td>
<td>3.37</td>
<td>15</td>
<td>33</td>
</tr>
<tr>
<td><strong>Class 4</strong></td>
<td>123</td>
<td>11.84</td>
<td>4.43</td>
<td>3</td>
<td>21</td>
</tr>
</tbody>
</table>

*Note.* Cut scores for Modified Angoff were 30.48 for *Minimally Proficient* and 40.66 for *Minimally Advanced.*
Table 10

Average Posterior Probabilities of Class Membership by Latent Class of the Championed Ordinal LCA Solution

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>0.910</td>
<td>0.090</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.068</td>
<td>0.880</td>
<td>0.053</td>
<td>0.000</td>
</tr>
<tr>
<td>Class 3</td>
<td>0.000</td>
<td>0.113</td>
<td>0.872</td>
<td>0.015</td>
</tr>
<tr>
<td>Class 4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.030</td>
<td>0.970</td>
</tr>
</tbody>
</table>

*Note.* Rows include students who were assigned to that particular latent class and columns represent the average posterior probability of class membership by all latent classes for students assigned to a particular class.
Table 11
Comparison of Student Classifications using Modified Angoff and 4-class Ordinal LCA

<table>
<thead>
<tr>
<th>Modified Angoff Category</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advanced</strong></td>
<td>694</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Proficient</strong></td>
<td>481</td>
<td>1443</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td><strong>Developing</strong></td>
<td>0</td>
<td>273</td>
<td>532</td>
<td>123</td>
</tr>
</tbody>
</table>

1175 (33.02%)  1720 (48.34%)  540 (15.18%)  123 (3.46%)  3558 (100%)

*Note.* Ordinal LCA classes are ranked from highest- (1) to lowest-achieving (4) on the ERIT-XA.
Table 12  
*Summary of Descriptive Statistics for Auxiliary Variables*

<table>
<thead>
<tr>
<th></th>
<th>SER Rank</th>
<th>SER Importance</th>
<th>SER Confidence</th>
<th>SER Engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>2896</td>
<td>2889</td>
<td>2886</td>
<td>2882</td>
</tr>
<tr>
<td>M</td>
<td>4.41</td>
<td>22.08</td>
<td>20.80</td>
<td>19.10</td>
</tr>
<tr>
<td>SD</td>
<td>2.41</td>
<td>3.26</td>
<td>3.08</td>
<td>4.18</td>
</tr>
<tr>
<td>Min</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Max</td>
<td>10</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Possible Range</td>
<td>1-10</td>
<td>5-25</td>
<td>5-25</td>
<td>5-25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Dosage1</th>
<th>Dosage3</th>
<th>Dosage4</th>
<th>Dosage5</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>914</td>
<td>913</td>
<td>913</td>
<td>907</td>
</tr>
<tr>
<td>M</td>
<td>3.36</td>
<td>3.27</td>
<td>3.06</td>
<td>2.40</td>
</tr>
<tr>
<td>SD</td>
<td>1.13</td>
<td>1.14</td>
<td>1.23</td>
<td>1.22</td>
</tr>
<tr>
<td>Min</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Max</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Possible Range</td>
<td>1-5</td>
<td>1-5</td>
<td>1-5</td>
<td>1-5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SOS Importance</th>
<th>SOS Effort</th>
<th>SAT Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>3252</td>
<td>3248</td>
<td>2916</td>
</tr>
<tr>
<td>M</td>
<td>14.77</td>
<td>19.70</td>
<td>567.19</td>
</tr>
<tr>
<td>SD</td>
<td>4.39</td>
<td>3.55</td>
<td>69.32</td>
</tr>
<tr>
<td>Min</td>
<td>5</td>
<td>5</td>
<td>260</td>
</tr>
<tr>
<td>Max</td>
<td>25</td>
<td>25</td>
<td>800</td>
</tr>
<tr>
<td>Possible Range</td>
<td>5-25</td>
<td>5-25</td>
<td>200-800</td>
</tr>
</tbody>
</table>

*Note. SER Rank refers to students’ average ranking of the importance of ethical reasoning to their life/career after graduation*
Table 13
Validity Analysis Results for Modified Angoff and 4-class Ordinal LCA

<table>
<thead>
<tr>
<th>Variable</th>
<th>Advanced</th>
<th>Proficient</th>
<th>Developing</th>
<th>df1</th>
<th>df2</th>
<th>F</th>
<th>p</th>
<th>$\omega^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SER Rank</td>
<td>4.26</td>
<td>4.39</td>
<td>4.56</td>
<td>2</td>
<td></td>
<td>2.55</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>SER Importance</td>
<td>22.64_a</td>
<td>22.28_a</td>
<td>21.24_b</td>
<td>2</td>
<td></td>
<td>37.01</td>
<td>&lt;0.001</td>
<td>0.02</td>
</tr>
<tr>
<td>SER Confidence</td>
<td>21.36_a</td>
<td>20.76_b</td>
<td>20.47_b</td>
<td>2</td>
<td></td>
<td>13.78</td>
<td>&lt;0.001</td>
<td>0.01</td>
</tr>
<tr>
<td>SER Engagement</td>
<td>19.56_a</td>
<td>19.18_a</td>
<td>18.62_b</td>
<td>2</td>
<td></td>
<td>8.71</td>
<td>&lt;0.001</td>
<td>0.01</td>
</tr>
<tr>
<td>Dosage1</td>
<td>3.60_a</td>
<td>3.42_a</td>
<td>3.04_b</td>
<td>2</td>
<td></td>
<td>15.74</td>
<td>&lt;0.001</td>
<td>0.03</td>
</tr>
<tr>
<td>Dosage3</td>
<td>3.38</td>
<td>3.26</td>
<td>3.17</td>
<td>2</td>
<td></td>
<td>1.90</td>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>Dosage4</td>
<td>2.96</td>
<td>3.05</td>
<td>3.14</td>
<td>2</td>
<td></td>
<td>1.17</td>
<td>0.31</td>
<td>0.00</td>
</tr>
<tr>
<td>DOSage5</td>
<td>2.28</td>
<td>2.41</td>
<td>2.47</td>
<td>2</td>
<td></td>
<td>1.49</td>
<td>0.23</td>
<td>0.00</td>
</tr>
<tr>
<td>SOS Import</td>
<td>15.04_a</td>
<td>14.81_ab</td>
<td>14.48_b</td>
<td>2</td>
<td></td>
<td>3.13</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>SOS Effort</td>
<td>20.59_a</td>
<td>19.96_b</td>
<td>18.53_c</td>
<td>2</td>
<td></td>
<td>73.18</td>
<td>&lt;0.001</td>
<td>0.04</td>
</tr>
<tr>
<td>SAT Reading</td>
<td>617.32_a</td>
<td>566.70_b</td>
<td>527.12_c</td>
<td>2</td>
<td></td>
<td>331.90</td>
<td>&lt;0.001</td>
<td>0.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
<th>df1</th>
<th>df2</th>
<th>F</th>
<th>p</th>
<th>$\omega^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SER Rank</td>
<td>4.30_a</td>
<td>4.39_ab</td>
<td>4.63_ab</td>
<td>4.89_b</td>
<td>3</td>
<td></td>
<td>3.24</td>
<td>0.021</td>
<td>0.00</td>
</tr>
<tr>
<td>SER Importance</td>
<td>22.58_a</td>
<td>22.20_a</td>
<td>21.39_b</td>
<td>18.54_c</td>
<td>3</td>
<td></td>
<td>53.99</td>
<td>&lt;0.001</td>
<td>0.05</td>
</tr>
<tr>
<td>SER Confidence</td>
<td>21.21_a</td>
<td>20.68_a</td>
<td>20.68_a</td>
<td>19.02_b</td>
<td>3</td>
<td></td>
<td>17.20</td>
<td>&lt;0.001</td>
<td>0.02</td>
</tr>
<tr>
<td>SER Engagement</td>
<td>19.62_a</td>
<td>19.00_a</td>
<td>18.77_a</td>
<td>17.15_b</td>
<td>3</td>
<td></td>
<td>13.14</td>
<td>&lt;0.001</td>
<td>0.01</td>
</tr>
<tr>
<td>Dosage1</td>
<td>3.61_a</td>
<td>3.30_ab</td>
<td>3.09_bc</td>
<td>2.72_c</td>
<td>3</td>
<td></td>
<td>11.50</td>
<td>&lt;0.001</td>
<td>0.03</td>
</tr>
<tr>
<td>Dosage3</td>
<td>3.36_a</td>
<td>3.25_a</td>
<td>3.23_a</td>
<td>2.48_b</td>
<td>3</td>
<td></td>
<td>4.84</td>
<td>0.002</td>
<td>0.01</td>
</tr>
<tr>
<td>Dosage4</td>
<td>3.17</td>
<td>3.04</td>
<td>3.04</td>
<td>2.88</td>
<td>3</td>
<td></td>
<td>0.64</td>
<td>0.59</td>
<td>0.00</td>
</tr>
<tr>
<td>Dosage5</td>
<td>2.60</td>
<td>2.42</td>
<td>2.31</td>
<td>2.08</td>
<td>3</td>
<td></td>
<td>2.51</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>SOS Import</td>
<td>14.93_a</td>
<td>14.83_a</td>
<td>14.43_a</td>
<td>13.33_b</td>
<td>3</td>
<td></td>
<td>5.76</td>
<td>0.001</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Mean (a)</td>
<td>Mean (b)</td>
<td>Mean (c)</td>
<td>df</td>
<td>F</td>
<td>Sig</td>
<td>Mean (a)</td>
<td>Mean (b)</td>
<td>Mean (c)</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>SOS Effort</td>
<td>20.42</td>
<td>19.87</td>
<td>18.68</td>
<td>15.23</td>
<td>3</td>
<td>3244</td>
<td>95.66</td>
<td>&lt;0.001</td>
<td>0.08</td>
</tr>
<tr>
<td>SAT Reading</td>
<td>606.91</td>
<td>553.68</td>
<td>526.34</td>
<td>535.06</td>
<td>3</td>
<td>2912</td>
<td>220.88</td>
<td>&lt;0.001</td>
<td>0.18</td>
</tr>
</tbody>
</table>

*Note.* Means are reported for each class/group. Different subscripts denote statistically significant differences between means. df = degrees of freedom.
Table 14
*Classifications by Participation in MC Intervention for Modified Angoff*

<table>
<thead>
<tr>
<th>Participated in MC Intervention</th>
<th>Modified Angoff Performance Category</th>
<th>N</th>
<th>%</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Developing</td>
<td>57</td>
<td>23.55%</td>
<td>24.25</td>
<td>5.45</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Proficient</td>
<td>117</td>
<td>48.35%</td>
<td>36.09</td>
<td>2.97</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>68</td>
<td>28.10%</td>
<td>43.44</td>
<td>2.00</td>
<td>41</td>
<td>48</td>
</tr>
<tr>
<td>No</td>
<td>Developing</td>
<td>871</td>
<td>26.27%</td>
<td>24.28</td>
<td>6.09</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Proficient</td>
<td>1815</td>
<td>54.73%</td>
<td>35.61</td>
<td>2.79</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>630</td>
<td>19.00%</td>
<td>43.08</td>
<td>1.83</td>
<td>41</td>
<td>50</td>
</tr>
</tbody>
</table>

*Note.* % column refers to the percentage of students categorized into that performance group within the specified MC subsample.
<table>
<thead>
<tr>
<th>Participated in MC Intervention</th>
<th>Ordinal LCA Class</th>
<th>N</th>
<th>%</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Class 1</td>
<td>110</td>
<td>45.45%</td>
<td>41.75</td>
<td>2.75</td>
<td>36</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Class 2</td>
<td>91</td>
<td>37.60%</td>
<td>33.51</td>
<td>2.98</td>
<td>28</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Class 3</td>
<td>32</td>
<td>13.22%</td>
<td>24.44</td>
<td>3.56</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Class 4</td>
<td>9</td>
<td>3.72%</td>
<td>15.00</td>
<td>3.91</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>No</td>
<td>Class 1</td>
<td>1065</td>
<td>32.12%</td>
<td>41.35</td>
<td>2.62</td>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Class 2</td>
<td>1629</td>
<td>49.13%</td>
<td>33.74</td>
<td>2.96</td>
<td>26</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Class 3</td>
<td>508</td>
<td>15.32%</td>
<td>24.80</td>
<td>3.35</td>
<td>15</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Class 4</td>
<td>114</td>
<td>3.44%</td>
<td>11.59</td>
<td>4.38</td>
<td>3</td>
<td>20</td>
</tr>
</tbody>
</table>

*Note. % column refers to the percentage of students categorized into that performance group within the specified MC subsample.*
Figure 1. Cluster profiles from GED writing skills test. Figure adapted from “Using cluster analysis to solve the problem of standard setting” by S. G. Sireci, 1995, Paper presented at the meeting of the American Psychological Association, New York, NY.
Figure 2. Cluster profiles from Grade 7 mathematics test. Figure adapted from “Using cluster analysis to facilitate standard setting” by S. G. Sireci, F. Robin, & T. Patelis, 1999, *Applied Measurement in Education, 12*, 301-325.
Figure 3. Two-class LCA profile plot of responses to a middle school math test. Figure adapted from “Using latent class analysis to set academic performance standards” by R. S. Brown, 2007, *Educational Assessment, 12*, 283-301.
Figure 4. Five-class MRM profile plot from reading subscale of a language proficiency test. Figure adapted from “Exploring levels of performance using the mixture Rasch model for standard setting “ by H. Jiao, R. W. Lissitz, G. Macready, S. Wang, & S. Liang, S., 2011, Psychological Test and Assessment Modeling, 53, 499-522.
Figure 5. Four-class example plot of non-parallel profiles.
1) Select a Standard-Setting Method
2) Choose a Panel & Design
3) Prepare Descriptions of Performance Categories
4) Train Panelists
5) Collect Item Ratings
6) Provide Feedback & Facilitate Discussion
7) Compile Panelist Ratings & Obtain Performance Standards
8) Conduct Panelist Evaluation
9) Compile Validity Evidence & Prepare Technical Documentation

Prior to Day 1
During Day 1
End of Day 1
After Study

Standard Setting Agenda

Ethical Reasoning Identification Test-XA (ERIT-XA)

8:00 – 8:30  Breakfast, Introductions, Orientation

8:30 – 9:15  Discussion of ERIT performance categories
             Discussion of standard setting (modified Angoff)

9:15 – 9:30  Break

9:30 – 9:45  Practice ratings

9:45 – 10:30 Round 1 ratings

10:30 – 10:45 Break

10:45 – 11:15 Discussion of Round 1 results

11:15 – 12:00 Round 2 ratings

12:00 – 12:15 Break

12:15 – 12:45 Discussion of Round 2 ratings

12:45 – 1:00  Wrap-up; Lunch

Figure 7. Standard setting agenda for the ERIT-XA.
Figure 8. Visual portrayal of “borderline” groups and performance categories for ERIT-XA standard setting.
Figure 9. Survey platform used by SMEs to provide item ratings for each borderline group.
Figure 10. Median item ratings provided by SMEs for each borderline group in Rounds 1 and 2.
Figure 11. Raw score distribution for subsample of students who took the ERIT-XA \((n = 1,840)\).
Figure 12. Percent-correct score distribution for subsample of students who took the ERIT-XA ($n = 1,840$).
Figure 13. Probability of correct response by item for the ERIT-XA for the 4-class general LCA solution.
Figure 14. Probability of correct response by item for the ERIT-XA for the 4-class ordinal LCA solution.
Figure 15. Path diagrams showing the connection between LCA (graphic A) and a factor model (graphic B).
Figure 16. Probability of correct response by item for the ERIT-XA for the one-factor model
Footnotes

1Demographic data were only available for the Assessment Day testing sessions (i.e., demographics were not collected during MC testing). Thus, the sample used for demographic data analysis only included students from the Assessment Day testing sessions used for analyses ($n = 3,316$).

2In an ordinal LCA, $C$ in the path diagram is an ordinal variable, whereas $C$ represents a nominal variable in a general LCA model.
Appendix

General LCA Profile Plots

Note. Probability of correct response by item of the ERIT-XA for the 1-class general LCA solution.
Note. Probability of correct response by item of the ERIT-XA for the 2-class general LCA solution.
Note. Probability of correct response by item of the ERIT-XA for the 3-class general LCA solution.
Note. Probability of correct response by item of the ERIT-XA for the 5-class general LCA solution.
References


