To God all the glory
ACKNOWLEDGMENTS

I thank my parents, my beautiful girlfriend, my teachers, my friends and Kanye West. I also thank John Oliver for inspiring this work.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................ 4

LIST OF TABLES ............................................................................................................. 6

LIST OF FIGURES ......................................................................................................... 7

ABSTRACT .................................................................................................................... 8

CHAPTER

1  INTRODUCTION ..................................................................................................... 10

2  LITERATURE REVIEW ......................................................................................... 16

   Validity of State-wide Assessments ....................................................................... 16
   Empirical Standard Setting ................................................................................. 19

3  METHODS ............................................................................................................ 39

   Data Simulation ................................................................................................. 40
   Model Comparisons ......................................................................................... 43
   Analysis ............................................................................................................ 49

4  RESULTS ............................................................................................................ 51

5  DISCUSSION ....................................................................................................... 58

   Data Simulation ............................................................................................... 58
   Model Comparisons ....................................................................................... 63

6  CONCLUSION .................................................................................................... 73

LIST OF REFERENCES ............................................................................................ 76

BIOGRAPHICAL SKETCH ...................................................................................... 80
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-1</td>
<td>FCAT 2.0 2013-2014 Item Parameters</td>
<td>42</td>
</tr>
<tr>
<td>4-1</td>
<td>Average Achievement Level Proportions</td>
<td>51</td>
</tr>
<tr>
<td>4-2</td>
<td>Standard Deviation of Achievement Level Proportions</td>
<td>53</td>
</tr>
<tr>
<td>4-3</td>
<td>Difference Between Achievement Level Proportions</td>
<td>54</td>
</tr>
<tr>
<td>4-4</td>
<td>Percentage of Agreement for Achievement Level assignment</td>
<td>56</td>
</tr>
<tr>
<td>5-1</td>
<td>Percent Agreement for 2013-2014 FCAT Reading Assessment</td>
<td>66</td>
</tr>
<tr>
<td>Figure</td>
<td>Achievement level assignments.</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------</td>
<td></td>
</tr>
<tr>
<td>4-1</td>
<td>57</td>
<td></td>
</tr>
</tbody>
</table>
The dynamic educational landscape is reflected in the consistent struggle to attain quality education for all. As the fight continues, appropriate scrutiny of current practices is needed for progress. The system's emphasis on accountability, with its primary metric being the standardized assessment, garners considerable criticism. Directing our attention to this facet of education highlights the standard setting process, which employed by state educational systems, is designed to attribute achievement levels to student assessment scores in order to make decisions throughout the framework of the educational system. Student learning gains, teacher effectiveness, graduation qualifications and school grades are just some of the metrics used in these decisions that rely on the appropriation of student’s scores on state-wide assessments to achievement levels.

Considerable perusal surrounding the use of standardized assessment scores and the legislative significance of the standardized assessment highlights this area of the educational system as an appropriate loci for improvement. This has traditionally been dominated by expert-based standard setting methods void of significant empirical
validation. The purpose of this study is to address the use of three clustering models as a tool for validation in the standard setting process. Comparing the resulting achievement level assignments of these three models revealed improved levels of accuracy in achievement level assignment and confirmed the results from previous study in a large-scale high-stakes environment.
CHAPTER 1
INTRODUCTION

One of the most consequential components of western society is our ability to learn and educate. This notion is emphasized by our educational system and its place in our culture. Every child in the US must attend school for a substantial portion of their upbringing. Compulsory Education Laws written by each state mandate primary education for our youth (Katz, 1976). Through the Elementary and Secondary Education Act of 1965 these laws are accompanied by mandated state funding to ensure that this developmental period is both a right and an opportunity (Jeffrey, 1978). However, merely requiring attendance and ensuring that every child has access to education does not guarantee that learning will happen, much less quality learning (Sciarra, 2015).

Our recent educational history would suggest that something is missing. The media and political entities have made comparisons between the educational system in the United States and the systems in other developed countries (OECD, 2015). These comparisons depict a US system that is failing to produce students with abilities that meet our culture’s standards (OECD, 2015). Yet, scrutiny of the American education system does not stop at comparisons to other countries around the world. A look into the achievements of all students suggested alarming gaps separating children from privileged backgrounds and those from less advantaged environments (Rothstein, 2004).

The efforts to close the achievement gap and the pursuit of educational merit led to paramount political actions to improve the educational environment in the US (West, 2003). Perhaps the most significant educational action in recent history was the No Child Left Behind Act of 2001 (NCLB). As it stood the NCLB Act was the culmination of
national educational reform engineered toward equal opportunity for all children. Cited in the opening statement was the reality that the federal government was not doing enough to promote quality education and prevent students from falling behind (No Child Left Behind Act of 2001 [NCLB], 2001). At the heart of the reform was increased accountability focused on measurable assessments (Kimmelman, 2006). These assessments provided the federal government, states, school districts and even citizens with the information to hold the educational system accountable for ensuring that all students, advantaged and disadvantaged, met high academic standards. There was a clear logic for this reform that stemmed from high standards and increased accountability through mandated assessments.

Demonstrating that the fight for quality education is far from over, the Every Child Achieves Act of 2015 (ECAA) replaced the NCLB with significant implications. The ECAA altered a principal element of the educational system that began with the NCLB Act in 2001. The ECAA made significant amendments to the accountability system relinquishing the confines of the NCLB Act and entrusting the states with near full responsibility in how academic accountability is measured (Every Child Achieves Act of 2015 [ECAA], 2015). The federal government still requires states to develop accountability systems and administer yearly standardized assessments with particular achievement for graduation, but the states are now able to determine how to administer these standardized assessments throughout the course of the year (ECAA, 2015).

Despite the educational reform of the last decade and a half, one element remained. Standardized assessments still exists as one of the most critical aspects of our accountability system in the US (Sciarra, 2015). Additionally standardized testing
garner considerable resistance and remains one of the most controversial and contested aspects of education (Oliver, 2015). The polarizing issues with standardized testing are the consideration for the uses of standardized test scores (Baker, 2010; Turniseed, 2015). As it stands today, the system of accountability rests primarily on the scores generated by students (Parsi, 2015). The students take the assessments and receive a scaled score that represents their performance. That score then determines their place in an achievement level. It is through the achievement level distinctions that meaningful inferences are made (Fla. Stat. § 1008.34). The achievement levels are the metric for which progress or the lack of progress is determined at a school, district or state level (Parsi, 2015).

The significance of the achievement level within the accountability system draws our attention to the processes that exist to create, monitor and utilize the achievement levels. In order to make judgments about the use of standardized test, it is first important to understand how meaning is assigned to the tests. The practice of standard setting exists for this purpose. The standard setting process for large scale high-stakes assessments has been dominated with what are described as test-centered methods (Jaeger, 1993; Stephenson, et al. 2000). These test-centered methods include the Angoff method (Angoff, 1971), the modified Angoff method (Hambleton & Plake, 1995; Impara & Plake, 1997; Taube, 1997), the bookmark method (Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001), the Nedelsky method (Nedelsky, 1954), the Ebel method (Ebel, 1972) and the Jaeger method (Jaeger, 1982) among others. There exist alternative methods that focus on the examinees rather than the tests. These methods rely principally on judgments about individual examinees’ performance. The
three commonly used methods are the contrasting groups method (Zieky & Livingston, 1977), the borderline group method (Zieky & Livingston, 1977), and the up and down method (Livingston & Zieky, 1982).

The most widely utilized standard setting methods are the modified Angoff method, the bookmark method, and the body of work method for assessments that contain items intended to measure performance or competency (Jiao, 2011). Although these methods differ in how they determine cut scores, both standpoints, the test-centered and examinee-centered methods, are based on the subjectivity of experts’ judgement. There can be limitations and even serious consequences when using only subjective judgment especially due the high-stakes implications of state and national assessments (Sireci, 1999). Although the expert-based methods are systematic and have storied histories, there are undoubtable errors in judgement for two reasons. Errors may always exist because the judgements rely entirely on subjective methods and because they do not employ any form of external or internal validation (Jiao, 2011).

Standardized assessments are the drivers of the accountability system and they are made meaningful through the use of achievement levels (Fla. Stat. § 1008.34). As standard setting is the method for defining the achievement levels it is a logical place to pass judgements in our effort to improve our educational system (Sireci, 1999; Brown, 2007; Jiao, 2011). Current standard setting models leave stakeholders without the ability to detect errors and are therefore vulnerable to make uninformed or poor decisions (Jiao, 2011). For this reason, researchers have extended the practice of standard setting to include objective methods that utilize item response data. These methods include latent class modeling (Brown, 2007; Luecht & DeChamplain, 1998;
Templin, Poggio, Irwin, & Henson, 2007), cluster analysis (Sireci, 1999; Sireci, Robin, & Patelis, 1999), mixture Rasch models (Jiao, 2011) and attribute hierarchy modeling (Sadesky & Gushta, 2004). Instead of expert judgement indicating group membership, model fit indices based on test data suggest the appropriate grouping. The latest work on this concept employs a mixture Rasch model (Jiao, 2011).

This body of literature suggests two applications for empirical modeling within the process of standard setting. The first is to use item response models as a means to validate the efforts of the judgment based procedures (Brown, 2007; Jiao, 2011). The second is to use these item response models to complete the standard setting process in its entirety rendering an alternative approach to the judgment based methods (Sireci, 1999; Brown, 2007; Jiao, 2011). Both of these goals have been explored through the studies cited above, however, there still remains many unanswered questions and areas for further research. Jiao (2011) cited the need for the results of the current work to be reevaluated with more realistic student assessment data and the value of having the various models compared in one study.

The intent of this study is to address these two limitations of the current body of literature from the context of using item based empirical models as a tool for the validation of traditional judgment based standard setting procedures. Real student data were used to simulate data for analysis. These data were used with the three models employed in the literature to date. The results from a univariate latent class model with total scores, a latent class analysis with item responses and a mixture Rasch model were compared to address one area of further exploration suggested by Jiao (2011). A comparison of the results from judgement based standard setting procedures and the
results of the empirical models can indicate how well these models align. Conclusions regarding these empirical models in terms of providing validity evidence were determined by the percentage of students assigned to each group. Additional conclusions about the empirical models were made through comparisons of the resulting class assignments for the three models.

To address the appropriateness of the criticism surrounding standardized assessments in state wide accountability systems, the first step is to be critical of the manner in which these tests are used. The uses of the assessments begin with the determination of achievement levels. This process is incomplete, but efforts have been made to improve that process. This study adds to the body of research supporting that effort with hopes of improving the paramount element of the accountability system, the standardized assessment.
CHAPTER 2
LITERATURE REVIEW

As mentioned in the introduction, there exists a body of research focused on using empirical models in the standard setting process. This study focuses on using these empirical models for validating the traditional judgment based procedures as support for the overall validity of the assessments used in state-wide accountability systems. This literature review will detail the aspects of empirical standard setting procedures that apply to validating current judgment based methods.

Validity of State-wide Assessments

Before discussing how the empirical models can provide validity evidence for the standard setting process, it is important to understand how the standard setting process relates to the overall validity of standardized assessments. Because the standard setting process gives meaning to student assessment scores and enables the scores to be used in a variety of forms this aspect of the accountability system plays a critical role in the validity of the standardized assessments (Fla. Stat. § 1008.34; Messick, 1987).

There is a void in the standard setting process that challenges the appropriateness of the uses of the standardized assessment in the system of accountability (Sireci, 1999; Brown, 2007; Jiao, 2011). Depicting the problems in the system of accountability frames how the standard setting process can be improved. The empirical item response based models are one way this improvement can be achieved. This logical notion begins at the concept of validity as it relates to standardized assessments.

The concept of validity is an ongoing area of study that has been extensively researched and explored in many contexts. Validity in terms of assessments was greatly influenced by the work of Cronbach (1955) and Messick (1980) among others.
Messick (1987) described validity as a judgment of the degree in which the empirical evidence and theory surrounding a measurement support the inferences and actions based on the results of that measurement. This means that for every measurement there must be theory and evidence to support actions made from the results of that measurement. With this notion, the manner in which an assessment is used becomes critical in judging its validity.

An anecdote from the state of Florida serves to illustrate the uses of standardized assessments enabling judgments to be made about the evidence to support those uses. Mandated by state of Florida law, students must earn a Level 2 (out of 5) on the Grade 3 English language arts assessment to be promoted to grade 4. A student must achieve a Level 3 (out of 5) on the Grade 10 English language arts assessment to graduate from high school. Students must also earn a Level 3 (out of 5) on the Algebra 1 End of Course assessment to graduate from high school. Courses whose subjects have a corresponding standardized assessment must include the results as 30% of the student’s course grade. The achievement and learning gains of students on the assessments are used to determine school grades, district grades, and school improvement ratings for alternative schools. Student performances on assessments determine whether schools require progress monitoring. A portion of a teacher’s evaluation are determined by the learning gains made by their students of the assessments that correspond to the subjects they teach ((Fla. Stat. § 1008.22).

Achievement and learning gains are determined by calculating the percentage of students who move from one achievement level to the next or by evaluating the changes in the number of students meeting proficiency levels.
School Grades; Fla. Stat. § 1012.34). Proficiency is defined by scoring Level 3 or higher on an assessment (Fla. Stat. § 1008.34). Also mandated by law, achievement levels are assigned to students based on their scaled assessment score which is an equated score calculated from the students raw score on the assessment (Florida Statewide Technical Report). Nearly all of the uses of the standardized assessments make use of the achievement level distinctions that are assigned to a student’s assessment score. The achievement level is intended to represent the student’s level of content mastery acquired in a particular subject (Fla. Stat. § 1008.34).

Standardized assessments are used to determine achievement gains which are calculated by improvements in achievement levels ultimately resulting in tangible implications for students, teachers, and school districts. In following the principles of validity by Messick (1987), the achievement level is highlighted as a critical element in validating the use of standardized assessments in the statewide accountability system. This distinction frames the judgement of the procedures in which achievement levels are determined. Noted in the introduction are the methods for which achievement levels are traditionally assigned and so too are the limitations of those methods (Angoff, 1971; Hambleton & Plake, 1995; Impara & Plake, 1997; Taube, 1997; Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001; Nedelsky, 1954; Ebel, 1972; Jaeger, 1982; Sireci, 1999; Brown, 2007; Jiao, 2011). Traditional standard setting models employ expert judgements to assign achievement levels to scores from standardized assessments. For example the modified Angoff method, one of the most widely used standard setting procedures, begins by selecting a group of experts to participate in the standard setting process. The experts develop a definition of a hypothetical minimally
competent practitioner (MCP) for each achievement level. The experts then review each item on the assessment considering how many of a group of 100 MCPs are likely to answer that question correctly. The Initial distinctions are discussed, and the experts reevaluate and change their decisions. This process is repeated for all item on the assessment until there is a cut point for each achievement (Angoff, 1971). However, the achievement levels must represent distinctly different groups of achievement to accurately serve their function within the accountability system (Fla. Stat. § 1008.34). Given that these traditional standard setting procedures do not utilize evidence to support that the achievement levels are in fact distinctly different, the resulting achievement level distinctions can be limited (USGAO, 1993; Sireci, 1999; Brown, 2007; Jiao, 2011).

**Empirical Standard Setting**

In order to address the void in the standard setting process, researchers looked for tools that accomplished the same goals as the standard setting process, but with less subjectivity (Brown, 2000; Luecht & DeChamplain, 1998; Templin, Poggio, Irwin, & Henson, 2007; Sireci, 1999, 2001; Sireci, Robin, & Patelis, 1999; Jiao, 2011; Sadesky & Gushta, 2004). There exists a collection of statistical models that cluster objects into discrete groupings (Anderberg, 1973; Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Rost, 1990; Dayton, 1991; Haertel, 1984, 1989; Luecht & DeChamplain, 1998). These models can obtain results that are theoretically similar to the judgment based methods of standard setting, but with additional advantages. Judgment based standard setting procedures generally begin with ordering students along a scaled continuum (Sireci, 1999). However, the results of standard setting are discrete groupings and there is support indicating that making discrete decisions based on
continuous scale is problematic (Dwyer, 1996; Jaeger, 1990; van der Linden, 1994). The grouping designations can prove to be inconsequential and arbitrary.

Sireci & Robin (1999) employed a cluster analysis to rectify the problem of using subjective judgments based on continuous scales in discrete standard setting decisions. Being the first to propose alternative methods for standard setting, Sireci & Robin (1999), employed a straightforward partitioning method that mimicked the goal of expert-based methods, separate students into achievement groups based on underlying achievement. This study sought to separate 818 seventh-grade students into qualitatively distinct groups. To do so, two clustering procedures were used based on empirical criteria. The first was a hierarchical cluster analysis (HCA) and the second was the K-means cluster analysis procedure. Two methods were used as the author explained, to achieve two slightly different goals. The HCA was executed to determine the number of groups that existed within the student population and then the K-means cluster analysis was used to place students into those groups ultimately determining the standards or partitioning criteria. The clustering techniques used by Sireci & Robin (1999) work by determining cluster centroids, a vector of means from the variables of interest, and calculating the distance individuals are away from these centroids. Each individual is assigned to the cluster that is the smallest distance away. The data that is imputed into these models were scores from non-mandated sections of the Connecticut Mastery Tests (CMT). The CMT is a statewide basic skills test that assesses a variety of skills in reading, writing, and mathematics. The CMT functions like many state-wide assessments as a means to group students into achievement levels. For the CMT there are three achievement levels; intervention, proficient, and excellence.
Sireci & Robin (1999) noted their discretion when determining the variables used to cluster the students within their population. The authors suggested that there were three options for selecting variables the first being all of the items on the assessment, the second being orthogonal factor scores derived from an item-even factor analysis, and third being sub scores calculated from the items within the major content areas of the assessment. Sireci & Robin (1999) cited the large number of items on the assessment they used for their clustering, the high intercorrelation among the content areas, and other non-specific recommendations based on previous research (Milligan, 1995; Sireci & Robin, 1996; Sneath, 1980) as their grounding for using total sub scores as the clustering variable. The decision to use total scores or sub scores is important as it defines one perspective in empirical standard setting procedures. As we will see, subsequent researchers will choose different methods.

After deciding upon how to select the variables in the cluster analysis, one of several options to calculate the distance between the centroids and the individuals was chosen. This distance metric is what ultimately indicates which cluster the individuals belong to. Sireci & Robin (1999) used the Euclidean distance for both clustering techniques to compute the distance. The distance is calculated with the formula:

\[
d_{ij} = \sqrt{\sum_{a=1}^{k} (x_{ia} - x_{ja})^2}
\]

\(d_{ij}\) is the distance between two objects \(i\) and \(j\) (for example between student \(i\) and \(j\) or between student \(i\) and cluster \(j\) ), \(x_{ia}\) is the score of student \(i\) on item \(a\), and \(k\) is the total number of items used to form the clusters.
Following the details of the variables used and the distance measure in the cluster analysis, Sireci & Robin (1999) describe the K-means clustering process. It begins with a designation of the number of variables used to cluster the students. In their study $K = 4$, meaning the centroid of each cluster was defined by a vector of the four means, and distances of the students to the cluster centroid were determined using Formula 2,1 stated above. Sireci & Robin (1999) use the SPSS software to conduct the K-means cluster analysis which employs the Anderberg's (1973) nearest centroid sorting method. The K-means process is iterative meaning that each individual can be assigned to different clusters until the conversion criteria is reached. In this case, conversion requires minimizing within-cluster distances and maximizing between-cluster distances. The resulting cluster solution was described as a Q-cluster solution (where Q represents the number of clusters) that resulted in the minimum value of the C-index for the HCA. The C-index (Dalrymple-Alford, 1970) describes the internal cohesion of the cluster solution, a metric that can inform the user which solution is appropriate. The authors note that the C-index performed well under simulation conditions when the actual number of clusters was known (Milligan, 1981).

For the HCA, Sireci & Robin (1999) use Ward’s (1963) method to cluster examinees. It was noted that Ward’s method was chosen because it produces relatively dense clusters. The authors cite that dense clusters were expected due to the relatively tight groupings of examinees. HCA begins by assigning each student to a separate cluster and sequentially merging students into groups until all students are in one group. The final step is to determine which cluster solution is most appropriate. The C-index
(Dalrymple-Alford, 1970) was used to help determine the appropriate number of clusters.

Aiding Sireci & Robin (1999) in determining the appropriate number of clusters were the mathematics grades earned by the students who participated in this study. Their final mathematics grades served as external criteria to validate the cluster solutions. The grades were reported on a 5-point scale ranging from F to A. Because there were three educational tracks within the participating school the grades were adjusted for the varying degrees of mathematics levels. Sireci & Robin (1999) admit that the adjustments to the final grades were not perfect; however, it provided improved information regarding the mathematics achievement, enabling the comparison to cluster solutions. The correlation between total scores on the test used in the cluster analysis and the recoded mathematics grades was 0.72.

To ultimately decide the most appropriate cluster solutions, a one-way analysis of variance (ANOVAs) was conducted on the K-means solutions. Cluster membership was the independent variable and school mathematics grade was the dependent variable. Sireci & Robin (1999) describe this analysis as an external evaluation of the cluster solutions because the total CMT scores were used to cluster the students rather than mathematics grades. In addition to the ANOVA, cross tabulations were computed between cluster membership and final mathematics grades. The criteria determined by Sireci & Robin (1999) were to deem cluster solutions valid if they grouped students in a manner that agreed with their final grades. Based on the results of the HCAs, two- through six-cluster K -means solutions were computed for both samples. In evaluating the stability of the cluster solutions, the two- and three-cluster solutions proved to be the
most consistent. The results from the one-way ANOVAs showed that clusters arising
from the two-, three-, and four-cluster solutions were significantly different from one
another in relation to the mathematics grades. Sireci & Robin (1999) noted that all pair
wise comparisons were not statistically significant for the five- and six-cluster solutions.
These results suggest that clustering the students beyond the four-cluster solution did
not provide any meaningful differences in student achievement. Sireci & Robin (1999)
stated given the ANOVA results, the two-, three-, and four-cluster solutions appear
viable, but after comparisons of the maximum and minimum scores between the lower
and higher proficiency groupings indicated that the three-cluster solution minimized
overlap differences among the clusters across replications yielding the three-cluster
solution the most appropriate. The three-cluster solution follows the number of
achievement levels that the CMT program uses in their scoring procedures. Sireci &
Robin (1999) found that their clustering method could serve as a complement to the
expert-based measures used in the CMT program.

As the first investigation into using empirical models in the standard setting
process, Sireci & Robin (1999) identified several implications and limitations of their
study. One of the elements of future research in this area would be the evaluation of the
stability of the cluster solution across samples. A second area of future research is the
exploration into further external validation metrics for the solutions. These future
applications can be achieved by using larger sample sizes and replicating the analyses
over several samples (Sireci & Robin, 1999). Additionally, Sireci & Robin (1999) note
the generalizability of the clustering approach needs to be addressed with different
types of assessments and score distributions.
Relating to the task of validating the traditional standard setting procedures, Sireci & Robin (1999) identified means to compare the results of the two approaches creating the opportunity to validate either perspective. By obtaining the achievement level assignments determined by the traditional judgment based method and comparing them to the achievement level designations from the cluster analysis, evidence for the appropriateness of the judgment based approach could be assessed. Granted that the cluster analysis determined the same number of clusters for the data that were used in the judgment based approach, using an empirical model could be useful in supporting the validity of the judgment based models (Sireci & Robin, 1999).

The next expansion of standard setting using empirical methods comes from Brown (2007). This study added to the discussion by designing a comparison between Expert-based procedures and empirical procedures. The empirical method used in this study was Latent Class Analysis. Brown (2007) cited a benefit of LCA that is similar to cluster analysis in that it does not assume that a continuous latent trait underlies student performance but rather groups exist within the data with distinct qualitative characteristics of which account for all differences in the data. The goals of the empirical procedure were to determine the degree to which a specified latent structure fits student performance data, determine the latent structure that best represents the relationships exhibited by the data, obtain item parameters estimates for each latent class, and lastly assign the responses to the most appropriate latent structure. The resulting latent structure was compared to the results of traditional Expert-based methods for the same assessments.
Brown’s study was unique in that the assessment used was a selection of items taken from a larger assessment. This meant that the Expert-based standard setting procedures needed to be conducted alongside the empirical LCA, in other words, there were not prior standards or cut scores from previous applications to compare the results from the LCA. This provided a comprehensive evaluation of the standard setting process, but also left many unanswered questions. There were no indication that the selection of items was subjected to the same rigorous reliability and validity checks that were inherent in the previous works given that they utilized assessments with established reliability and validity evidence. Despite the limitation in this regard, the use of LCA as the empirical standard setting method is a logical extension of the cluster analysis, as it incorporates item level responses into the grouping algorithms (Dayton 1991; Haertel, 1984, 1989; Luecht & DeChamplain, 1998).

Brown (2007) pointed out that LCA is generally used to determine the number of latent classes within a sample, identify the proportion of the sample within each latent class, and elicit conditional item probabilities within each class and also to make predictions regarding class membership based on response patterns. These inherent uses fit well with the logic and purpose of standard setting making LCA an appropriate choice as an empirical method. Brown (2007) further explains in detail the mechanisms that allow LCA to accomplish these tasks. The author noted that the conditional probability of a response pattern for a particular latent class was determined with the following formula:

\[
P(y_i | c) = \prod_{j=1}^{k} (\alpha_{jc})^{y_{ij}} \cdot (1 - \alpha_{jc})^{1-y_{ij}}
\]  

(2-2)
The unconditional probability of a given response pattern was obtained by using a weighted sum across all latent classes and the following formula:

\[ P(y_i) = \sum_{x=1}^{c} \theta_c \left[ \prod_{j=1}^{k} (\alpha_{jc})^{y_ij} \cdot (1 - \alpha_{jc})^{1-y_ij} \right] \]  

Bayes' Theorem was used to determine the probability of membership in a latent class given a particular response pattern. The Bayes' Theorem formula:

\[ P(c | y_i) = \frac{P(y_i | c) \cdot P(c)}{\sum_{c=1}^{c} P(y_i | c) \cdot P(c)} \]  

Brown (2007) designed this study to compare Expert-based methods with empirical alternatives. This was accomplished by employing educational experts to evaluate the section of items mentioned above using the modified Angoff method and a profile rating procedure. The same items were then evaluated using a LCA. The results of the methods were compared to assess the effectiveness of the LCA in a standard setting procedure. The assessment was comprised of ten multiple choice questions and two performance assessments from the probability and data representation section of the Third International Mathematics and Science Study designed for seventh and eighth graders. These items ranged in difficulty from relatively easy (\( p = .85 \)) to difficult (\( p = .41 \)), where \( p \) is the proportion of correct responses. There were two additional performance assessment items that aimed to measure a similar concept of probability. The population of individuals who participated in the study was a group of seventh- and eighth-graders from schools within a close proximity. There were 88 male students and 93 female students all of which were currently enrolled in mathematics courses.

As stated above, there did not exist a specific standard setting framework for this specific selection of items from any perspective. Therefore, Brown (2007) arranged an
Expert-based standard setting workshop in which 89 mathematics teachers from middle and high schools in the surrounding areas participated in two standard setting sessions as the raters of student achievement. The methods that were employed were the modified Angoff method and a profile rating procedure. With the modified Angoff method, the raters were asked to determine 36 different item ratings based on three different classifications for each of the 12 items. The classifications were basic, proficient and advanced. The raters were given an opportunity to modify their ratings after seeing item characteristics like difficulty and discrimination. Final item ratings were decided upon by assessing reliability metrics between raters. To determine performance standards, Brown (2007) decided to use the median percentage correct rating for the proficient category yielding 65.83% as the proficient cut-score for the Angoff method.

The profile rating procedure was employed as an alternative expert-based standard setting procedure and provided a slightly varied perspective than that of the Angoff method. The profile rating task consisted of the raters assigning 75% of the possible scoring profiles to categories of performance. The four categories were similar to the categories used for the Angoff procedure but incorporated one additional level. The performance categories included below basic, basic, proficient and advanced. Each rated assessed 120 unique scoring profiles with an additional 20 randomly selected and repeated profiles for a total 140 scoring profiles. The authors used two methods of determining the reliability and consistency of rating. One approach was to assess the percentage of times a rater would assign the same category to the same profile that was one of the randomly selected and repeated profiles. The second measure of consistency used the same sample of repeated scoring profiles, but recoded each item
to be a dichotomous variable where 0 indicated a response that was below basic or basic and a 1 indicated proficient or advanced. The percentage of agreeing determinations for the proficient profiles was calculated. The results of these consistency measures showed that 73.15% of the repeated ratings receiving the same categorical assignment and 87.6% of the repeated ratings had the same proficiency ratings (Brown, 2007).

After completing the Expert-based standard setting procedures, the focus shifted to the four goals of the LCA. A series of Latent class models were fit to the data using MPlus software (Muthén and Muthén, 1998). A total of 6 latent class models were fit to the data. Three of the models used each item as a binary response and three of the models used the items as continuous indicators. For the binary items each multiple choice question was score as either correct (1) or incorrect (0). The performance assessment item’s four achievement levels were dichotomized at the mid-point meaning a 1 was recorded for an achievement level of a 1 or 2 and a 0 was recorded for an achievement level of 3 or 4. For each variable type, binary and continuous, the models were fit to the data, one with a single class, one with two classes and one with three classes.

The resulting LCA models were assessed for model fit. To determine the model with best fit, several indexes were used that included the loglikelihood value relative to the number of parameters, the Akaike information criterion (Akaike, 1987), and the Bayesian information criterion (Schwartz, 1978). For the model using binary indicators, the AIC values were 2393.62, 2301.19, and 2307.25 for the 1-, 2- and 3-class models respectively. The AIC for the model using continuous indicators were 1570.27, 1529.42
and 1535.28 for the 1-, 2- and 3-class models respectively. For both variable type, the 2-class model had the lowest AIC value, supporting that this model fit the best. The BIC values for the model with binary indicators were 2432.84, 2382.89 and 2424.90 for the 1-, 2- and 3-class models respectively. For the model with continuous indicators, the BIC values were 1589.46, 1561.40 and 1580.06 for the 1-, 2- and 3-class models respectively. Similarly with the AIC indices, the model with 2-classes fit the best for both variable types. The loglikelihoods for the model with binary indicators were -1184.81, -1125.60 and -1117.63 for the 1-, 2- and 3-class models respectively. The loglikelihood increases from the 1-class model to the 2-class model significantly but only increases nominally from the 2-class model to the 3-class model. The loglikelihoods for the model with continuous indicators were -779.14, -754.71 and -753.64 for the 1-, 2- and 3-class models respectively. The loglikelihood increases from the 1-class model to the 2-class model significantly but only increases nominally from the 2-class model to the 3-class model. The loglikelihood values from both model types suggest that the 2-class model fit the most appropriately. Based on the fit indices provided in the study, Brown (2007) concluded that the 2-class models fit better than the 1-class or 3-class models.

In addition to the fit indices, Brown (2007) discussed the parameter estimates and the proportions of students in each class for each of the 6 models. The parameters that were estimated were item discrimination for the binary models and the mean score for the continuous models are depicted for each of the models with different numbers of classes. These estimates, along with the proportions of students in each class, add context to the conclusions made regarding the fit indices by providing an observable effect of the different class assignments between models. One of the most important
distinctions obtained from the parameter estimates was a discrepancy between the two 2-class models. The item level data in the binary model suggested there were a larger proportion of students characterized as high achievers, while the model using continuous indicators suggested that there were a far lower percentage of students characterized by high achievers. This disparity was addressed when the results from the empirical methods were compared to the results from the judgement based methods.

To make comparisons between methods, Brown (2007) utilized a 2x2 cross-tabulation sequence comparing the Angoff methods, profile rating methods, the binary LCA and the continuous LCA. The element that was compared between models was whether or not the different methods made the same distinction for each student of whether they met or did not meet the proficiency standards. Given that the LCA model contained a different number of classes than was used in the judgment based models, the proportion of class membership could not be compared between the empirical and judgment based models. Nonetheless, the comparison between the numbers of students deemed to be proficient by each model expressed by the percentage of agreement between models was quite high. The Angoff method and profile rating method were in agreement for 85.7% of the students. The Angoff and binary LCA method were in congruence for 92.2% of the student. The Angoff and the continuous LCA made the same determination for 66.2% of the students. The profile rating method and the binary LCA method were in agreement for 77.1% of the students, while the profile rating method and continuous LCA method were in congruence for 87.2% of the students. Lastly the binary LCA method and the continuous LCA made the same
judgments for 61.9% of the students. These values reveal that there were methods that shared very high levels of consistency in determining whether a student was determined proficient or not. However, there were two instances where the methods did not agree at a high level. Granted 4 of the 6 methods produced consistent determination for greater than 75% of the student and 3 of the 6 were above 85%. Given this information Brown (2007) made the conclusion that the results were sufficient to support that the various methods are concurrent in their ability to distinguish proficient student achievement.

After presenting the results and coming to a conclusion, Brown (2007) provided a discussion about the results in the larger context of standard setting. The discussion began with a portrayal of the limitation of the study. The limitation of the study according to Brown (2007) exists with the data. The data used was a small sample of items from one subject for one grade level. Brown (2007) called for the use of wide-scale norm-referenced assessments like the ones that are used heavily across the country in state level standardized assessments. Using more realistic data will add grounding to the use of these methods, as well as, an element of practicality. Following the brief anecdote about the limitation of this study, Brown (2007) highlights several implications of the study. First addressed was the disparity of the number of classes suggested but the latent class models and the number of classes used in the judgment based methods. This brings up the notion that the data may reveal a different number of groupings than is used with expert-based standard setting procedures and attempting to characterize students with more groupings than is supported by the data may be inappropriate. This implication provides another example of the effects of not having empirical validation.
measures for the standard setting procedures employed in the educational system. Alternatively, the results regarding the comparison between models and the ability to determine proficient achievement groups are promising. The comparison results are examples of how that validation process could be successful and is encouraging to other researchers to expand upon this study.

The last implication noted by Brown (2007) was described as the most important. The author supported prior claims that different judgment based standard setting procedures provide contrasting results. However, this study found that the Angoff method and the profile rating method agree for 85.2% of the students. The author suggests that this is an important finding; however, it is important to note the scope of the study when addressing the importance of the results. Most standard setting procedures in high stakes assessments are on a different scale than this study. The total sample of students was 181, whereas state level assessments are administered to groups well over 100,000. This implication may be interesting and important in some context, but the limitation in scope of the study diminished the impact for standard setting for large scale assessments.

One area that was not addressed was the disparity between the latent class models. This result has implications for the use of empirical measures in standard setting, but was not directly addressed in Brown (2007). As we will see with subsequent studies, the need for a comparison between empirical standard setting methods is beginning to take shape in the literature. Sireci (1999) employed a cluster analysis and Brown (2007) utilized a latent class model and subsequent work used other alternative models. No study has placed focus on empirical model comparison.
Although no study has yet compared empirical standard setting model performance, there was a study following Brown (2007) that extended the concept of empirical standard setting models. The work of Jiao (2011) is the most recent and significant contribution to empirical standard setting models. Jiao (2011) developed a framework for empirical standard setting that extends the most current explorations at that time. The study included a mixture IRT model that is able to classify individuals into groups and estimate latent ability simultaneously. The previous methods explained thus far have not been able to identify latent ability, rather only identity group membership. Much like the aforementioned studies, Jiao (2011) employed the mixture Rasch model to satisfy two goals. The first goal was to evaluate the performance of the mixture Rasch model as a tool to validate the performance proficiency levels set by the current Expert-based standard setting procedures. The second goal was to demonstrate how to determine performance cut scores based on the result of the mixture Rasch model.

Satisfying these two goals will contribute to the overarching exploration toward more efficient and justifiable standard setting procedures that reduce the level of subjectivity that exists in the current common practice of standard setting.

Jiao (2011) introduces the mixture Rash model as a tool developed by Kelderman and Macready (1990), Mislevy and Verhelst (1990) and Rost (1990) used to test data with more than one latent group. They achieved this by combining a Rasch measurement model (Rasch, 1960) and a latent class model. The Rasch model assumes that a continuous latent trait underlies the data, and the latent class model assumes there is a latent class membership that exists within the data. The mixture model assumes that examinees exist within latent groups or classes, but also there are
unique latent traits that exist within each of those groups or classes. It is through the mixing of these two models that the estimation of class membership and latent ability is possible.

As the mixture model contains elements from two distinct models of their own, each examinee is characterized by two latent variables, a continuous quantitative variable that reflects the examinees latent ability and a categorical qualitative variable fixing them to a particular group.

The mixture Rasch model looks to determine the probability of an individual correctly responding to an item with the following formula:

\[
p_{ji} = \frac{1}{1 + \exp\left[-(\theta_{ij} - b_{ig})\right]} \tag{2-4}
\]

The probability is estimated for each person \( j \) for a specific item \( i \) conditional on their class membership \( g \). The estimation also includes the item difficulty \( b_{ig} \) for the given class and the individual’s ability level \( \theta_{ij} \). The examinees’ latent group membership is determined by comparing the posterior probability of that particular examinee being in each latent group. The examinees are assigned to the late group that has the highest probability. Understanding how to assign group membership is vital to using this mixture model in the context of standard setting validation. Jiao (2011) cites the goal of all standard setting procedures as the identification of distinct groups in terms of the latent ability using the information in the item responses. The mixture Rasch model estimates the probability of group membership by incorporating the latent ability \( \Theta_{ig} \) making it particularly suited for the standard setting process.

To begin the exploration into using mixture Rasch models for standard setting, Jiao (2011) simulated data based on the standard setting results from a large-scale
assessment using the bookmark method. The data were to resemble a reading test with five proficiency levels. The proficiency level groupings were separations defined by different levels of theta or ability level; -1.8, -0.6, 0.96, and 1.68. These grouping characteristics yielded groups with the following proportions, 6% for group 1, 13.2% for hours 2, 52.6% for group 3, 22.3% for group 4, and 5.9% for group 5. The data were simulated with the assumption that the examinees within each proficiency level were from distinct normal distributions. Data were simulated for a total of 10,000 examinees yielding 600 examinees in group 1, 1300 examinees in group 2, 5300 examinees in group 3, 2200 examinees in group 4 and 600 examinees in group 5. The data that was simulated were item responses for all 10,000 examinees across 40 items, a commonly used test length in large-scale assessments.

Following the data simulation, seven Rasch models were fitted to the response data with increasing numbers of latent classes (1, 2, 3, 4, 5, 6, 7) using the mldtm software. Model fit was also evaluated using Akaike’s (1974) information criterion (AIC), Schwarz’s (1978) Bayesian information criterion (BIC), and deviance obtained from the mldtm software. The fit indices were not consistent in their support for the model with the best fit. The deviance and the AIC suggested that the 5-class model was the most appropriate model while the BIC supported the 3-class model as the model with the best fit. Because of the discrepancy, Jiao (2011) examined the estimated mean, standard deviation, and mixing proportions for the 5-class and the 3-class mixture Rasch models. The proportions of examinees assigned to the five classes respectively were, 6.26%, 12.57%, 52.86%, 22.11% and 6.19%. The estimations made by the model corresponded very closely to the proportions of students that existed in the simulated
data. Jiao (2011) noted that the results from the study support using mixture Rasch modeling for comparing multiple standard setting methods. This comparison can be a tool in the validation process.

Despite the success of the Rasch model, Jiao (2011) did highlight some important limitations of this study. The first source of limitations is in the data simulation procedure. According to Jiao (2011) the data was generated using relatively ideal conditions. The true mean and standard deviation were simulated with the assumption that the intersecting point for two adjacent distributions was two standard deviations away from the mean. Jiao (2011) noted that real test data would more than likely not be as ideally distributed and for this reason calls for more research with data that reflects more realistic testing distributions. Another extension of this study mentioned by Jiao (2011) highlights an area that was left unexplored by this study. Jiao (2011) indicated that this study did not incorporate model comparisons in that the mixture Rasch model was not compared with any other methods for standard setting, empirical or judgement based.

As the most contemporary exploration into empirical standard setting models, Jiao (2011) presents promising results while leaving many unanswered questions. The purpose of the current study is to address the limitations in the standard setting process employed by the accountability system in primary and secondary education in the US. As mentioned previously, the shortcomings of the standard setting process exist due to the lack of empirical validation in the standard setting process. The seminal studies highlighted above indicate the efforts to consummate the standard setting process by addressing the lack of empirical validation. However, Jiao (2011) suggested several
directives that remain to be explored. The study engages two of these notions. The first
one is the need for more realistic test data. Previous studies have employed data that
demonstrates the potential utility of these empirical processes, but the data used lacks
practical relevance (Jiao, 2011). The second notion explored in this study is the need to
compare different models used throughout the two decades of work in empirical
standard setting procedures. The myriad of models used to address the lack of
objectivity in the standard setting process have not been compared in one study to date.
This study addressed this gap in the literature. In addition to comparing the empirical
models, this study sought to cross-validate the results of the model comparisons with
results from the Expert-based standard setting procedures substantiating more
evidence for the use of empirical procedures as a validity tool in the standard setting
process.
CHAPTER 3
METHODS

The review of literature with regards to empirical standard setting procedures suggests many areas of continued study. The scope of this study addressed two of these areas. The first avenue of continued research addressed in the study is the need for the incorporation of more realistic data into the evaluation of empirical standard setting procedures. The second avenue is the need to compare multiple empirical standard setting procedures in one study. The first empirical was carried out through careful, yet practical data simulation. Details of the data simulation procedure are detailed below. The second empirical was addressed by executing multiple empirical models using the same data and comparing each model’s performance against the other empirical models and more importantly against the results of judgment based standard setting procedures. Model performance was judged based on the ability of the empirical models to partition the data into the appropriate class proportions determined by the judgment based methods. The decision to judge the performance of the empirical models was developed from one of two perspectives regarding the use of empirical models in the standard setting process.

The literature suggests that there are two distinct goals for empirical standard setting procedures in the general sense. One purpose is to use these empirical models as metric to validate or work in coordination with purely judgment based standard setting procedures. The second purpose is to use empirical models as an alternative standard setting procedure that would allow practitioners to accomplish the same goals of standard setting without the use of a purely judgment based methodology. The second purpose is explored quite extensively in the literature; however, the scope of this
study is to address the first purpose, using empirical modeling as a validation tool in congruence with a judgment based method. This was important to distinguish, as it shaped how the performances of these empirical models were judged. The purpose for empirical standard setting procedures that is not addressed by this study would extend our definition of model performance further to make judgments about where to define cut-scores between achievement levels. There is a subjective component to that decision making that is not in the scope of this study however should be addressed in future research. Narrowing the focus of the study defines the specific elements of model performance that were being judged for each empirical model.

Data Simulation

As addressed above, using data that more accurately resembles true student assessment data is one of the areas of further study for the field of empirical standard setting. With this said, using real student assessment data would be the most appropriate and valuable, however, given the high stakes nature of state wide assessments, actual student assessment scores, like item responses are very hard to obtain and in some cases protected by law (Jiao, 2011). Consolations were made in effort to remain within the scope of practicality while obtaining data that resembles true assessment data as closely as possible.

This study utilized simulated item response data for all three empirical models. The item response data used in this study were simulated in R software platform using parameters taken from actual reports of a particular set of state wide assessments. The parameters were also gathered from the state of Florida FCAT Yearbook, a report summarizing the state exam. These reports contain many test statistics that were used to simulate item response data that represent a true sample of student’s item responses
as closely as possible. The availability of the item characteristics of the data made the state of Florida assessments an appropriate example to use in this study. The simulation models generated a sample of 45 items for a total of 180,000 students. The number of items used by the state of Florida for their state wide assessments is 45 (FCAT Yearbook, 2014). The number of students was chosen to fit within the range of students who took the state of Florida FCAT Reading assessment. For the academic year 2013-2014, the state of Florida tested students in reading and mathematics. The number of students per grade level for the mathematics assessments ranged from 158,089 to 202,472 and the number of students per grade level for the reading assessments ranged from 166,447 to 203,184 (FCAT Yearbook, 2014).

The data was simulated according to a three parameter logistic model (3PL). The 3PL model is defined by the following formula:

\[ p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}} \]  

(3-1)

\( \theta \) represents the person’s ability and the item parameters \( a_i \), \( b_i \), and \( c_i \) represent the discrimination, difficulty and guessing parameters respectively (Embretson, 2013). The item parameters used to simulate the data for this study were obtained through the FCAT Yearbook report displayed in Table 3-1. (FCAT Yearbook, 2014).

After determining the number of students to simulate and the number of items, it was necessary to determine the item level parameters that would be used to generate the data. The assessment reports from the state of Florida contained numerous statistics that were incorporated into the data simulation algorithm. Item level difficulty, discrimination, and guessing parameters were listed for each grade level and subject in the state assessment report. The ranges of mean raw scores, standard deviations, item
discrimination, item difficulty and item guessing were all within ranges that suggested there were no significant differences between the grade levels or subjects. This enabled the use of only one sample of student scores as a sufficient test sample given the fact that repeating the models based on a different grade level or subject would not provide any additional meaningful insight. The item parameters that were used to simulate the data included a vector of item difficult values with the following quartile distribution, -2.340, -0.951, -0.514, -0.075, 0.740 corresponding to the minimum, 25th percentile, median, 75th percentile and maximum respectively. A vector of item discriminations, 0.532, 0.708, 0.936, 1.136, 1.742 corresponding to the minimum, 25th percentile, median, 75th percentile and maximum respectively was also used. Guessing parameters, 0.005, 0.091, 0.149, 0.190, 0.342, corresponding to the minimum, 25th percentile, median, 75th percentile and maximum were included as well. The maximum and minimum theta values were set to 3 and -3 respectively with the average theta being 0 with a standard deviation of 1. The values used in the data simulation model were taken from the detailed report for the FCAT for the 2013-2014 school year.

Table 3-1. FCAT 2.0 2013-2014 Item Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Min</th>
<th>25th Percentile</th>
<th>Median</th>
<th>75th Percentile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.532</td>
<td>0.708</td>
<td>0.936</td>
<td>1.136</td>
<td>1.742</td>
</tr>
<tr>
<td>B</td>
<td>-2.340</td>
<td>-0.951</td>
<td>-0.514</td>
<td>-0.075</td>
<td>0.740</td>
</tr>
<tr>
<td>C</td>
<td>0.005</td>
<td>0.091</td>
<td>0.149</td>
<td>0.190</td>
<td>0.342</td>
</tr>
</tbody>
</table>

Following a determination of the item specific parameters was the generation of data that fit the achievement profile suggested by the detailed report. For the 2013-2014 school year, the FCAT Reading assessment had achievement level proportions of
18.69% for Level 1, 24.42% for Level 2, 23.23% for Level 3, 23.46% for Level 4 and 10.21% for Level 5 (FCAT Yearbook, 2014). These parameters were used to generate an ability level distribution that would simulate data that mimicked this profile. The simulation of the test data was replicated 100 times and each replication was subjected to three clustering algorithms.

**Model Comparisons**

The outstanding literature in the field of empirical standard setting suggests the need to compare in one study the different approaches that have been used in empirical standard setting research efforts to date. Currently there is a void in the literature that leaves practitioners with no context explaining which empirical standard setting models can provide the most accurate partitioning of students (Jiao, 2011; Brown, 2007; Sireci, 1999). As mentioned previously, this served as the second empirical of this study. Completing this empirical involved comparing three models using identical data and evaluating the model’s performance as a validity metric for the standard setting process at large.

To judge a model’s performance as a tool for validation within standard setting, one of the simplest metrics would be to assess whether the empirical models partitioned the students into the same class proportions as the purely judgment based methods. The empirical models that were selected for this comparison were a latent class model using total scores to cluster students similar to the model used in Sireci (1999), a latent class model using item responses like the model employed in Brown (2007), and a mixture Rach model like the model used in Jiao (2011). Utilizing empirical latent trait or latent class models to partition students into achievement level as a means to validate traditional standard setting procedures follows the subsequent logic. Student
achievement is theoretically classified within a population of students to be a discrete characteristic. In some cases students can be characterized as being proficient or not proficient. In other cases like with the state of Florida state-wide assessments students can be classified into more groups such as, inadequate, below satisfactory, satisfactory, above satisfactory and mastery (Fla. Stat. § 1008.34) Although there are varying perspectives in which to classify students, the underlying logic is that discrete groupings exists. The models that were utilized in previous studies sought to partition students into groupings that represented these underlying discrete achievement levels based on certain criteria. Granted that criteria changed from model to model, but the underlying principles were the same throughout each model. The empirical of this study is to assess the performance of these models using the same data, an area void in the field of study. Each model is described to facilitate replication and support transparency in the standard setting process. The models are not described in a particularly meaningful order, but rather presented in the order which they were arrived in the evolution of empirical standard setting procedures.

The first model that used in the comparison was a univariate mixture model. The single viable used to cluster the hypothetical students was the total scores amassed from the simulated items. This model is comparable to the strategy employed by Sireci (1999). Although that study used a cluster analysis, the use of total scores as the clustering variable was replicated in this model. The univariate MM was executed using the Mplus software (Muthén & Muthén). The mixture model estimated mixture proportions with the formula:
The likelihood \( L_i \) for the ith observation within the population is calculated by substituting the value of \( f(x_i) \) in the following formula:

\[
f(x_i) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) e^{\left(-\frac{1}{2}\right) \left(\frac{(x_i - \mu)^2}{\sigma^2}\right)}
\]

The mixture proportions are calculated by maximizing \( \Psi \) in Equation 3-2 (Gagné, 2006).
shares a similar theoretical grounding as the multivariate LCA, but employs a different approach to determine the underlying groups that exist within a sample of data. Generally speaking a LCA model aims to partition a group of data into smaller subgroups based on underlying latent characteristics. The model will generate the proportion of students in each subgroup and the conditional item probabilities within each subgroup. The LCA model can also predict which subgroup a particular response pattern would fall under.

As stated with the univariate MM, the same model configurations used in earlier studies were employed in this study to reexamine the results using more realistic data and in order to compare the results to the other empirical measures. The data that was used in the LCA was the simulated item responses detailed in the Data Accumulation section. Brown (2007) cites the work of Dayton (1991) as the guidelines for executing the LCA; the same specifications were employed in this study. The LCA was executed using the Mplus software (Muthén & Muthén). These conditions state that given, \( y_i = (y_{ij}) \) be the vector of 1/0 responses by the \( i^{th} \) respondent \( (i = 1,\ldots, n) \) to the \( k \) items \( (j = 1,\ldots,k) \), \( \alpha_{jc} \) = the conditional item probability of item \( j \) in latent class \( c \), where \( (c = 1,\ldots,C) \), \( \Theta_c \) = the proportion of respondents in each latent class. The sum of these proportions across all classes must sum to 1.

These conditions state that given; \( y_i = (y_{ij}) \) be the vector of 1/0 responses by the \( i^{th} \) respondent \( (i = 1,\ldots, n) \) to the \( k \) items \( (j = 1,\ldots,k) \), \( \alpha_{jc} \) = the conditional item probability of item \( j \) in latent class \( c \), where \( (c = 1,\ldots,C) \), \( \Theta_c \) = the proportion of respondents in each latent class. The sum of these proportions across all classes must sum to 1.
The conditional probability of a response pattern given a particular latent classification is estimated using the following product-multinomial:

\[ P(y_i|c) = \prod_{j=1}^{k} (\alpha_{jc})^{y_{ij}} \cdot (1 - \alpha_{jc})^{1-y_{ij}} \] (3-4)

The unconditional probability of a given response pattern is estimated by using a weighted sum (weighted by the corresponding latent class proportion) across all latent classes:

\[ P(y_i) = \sum_{x=1}^{c} \theta_{cx} \left[ \prod_{q=1}^{k} (\alpha_{qc})^{y_{iq}} \cdot (1 - \alpha_{qc})^{1-y_{iq}} \right] \] (3-5)

The probability of membership in a latent class given a particular response pattern is then estimated using Bayes' Theorem:

\[ P(c|y_i) = \frac{P(y_i|c) \cdot P(c)}{\sum_{x=1}^{c} P(y_i|c) \cdot P(c)} \] (3-6)

These functions of the LCA make it possible to use a sample of student responses to determine the extent to which a specified latent structure fits the selected data, identify the latent structure that is most appropriately represented in the data, acquire estimates of item parameters for each latent class and predict the latent class membership based on a given response pattern. Therefore, the LCA model can satisfy the empirical of validating judgment based standard setting procedures by assigning examinees to latent groups which can be compared to the assignments determined by the judgement based methods and other empirical models.

Due to the data being simulated, the number of underlying classes within the data was known a priori. This limited the utility of the LCA to providing an achievement
level profile for each iteration of simulated data. The resulting class proportions and the corresponding item responses and total scores were collected for each iteration.

The third model that was compared in this study was the mixture Rasch model which was introduced into standard setting by Jiao (2011). Jiao (2011) cites Kelderman and Macready (1990), Mislevy and Verhelst (1990) and Rost (1990) as the pioneers of the model which are typically employed to model the test data with more than one latent population. This is possible by combining a Rasch IRT model (Rasch, 1960), which measures a continuous latent trait that underlies the achievement of examinees, and a latent class model, like the LCA described above. The combination of the Rasch model and the latent class model enables the simultaneous estimation of a continuous latent ability and latent group assignment. This simultaneous estimation of ability and group membership made the mixture Rasch model a logical extension in the study of empirical standard setting. Incorporating this model into this study involved utilizing the model employed by Jiao (2011).

The mixture Rasch model was executed with the Mplus software (Muthén & Muthén). The data explored by this model was the same simulated item responses that were used with the LCA model. The parameters of the mixture Rasch model are detailed here. The probability of a correct response for person j to item i conditional on the person’s latent class membership g is expressed in the following equation:

\[ p_{jig} = \frac{1}{1 + exp[-(\theta_{jg} - b_{ig})]} \]  

(3-7)

where \( p_{jig} \) is the probability that the \( j^{th} \) examinee will answer the \( i^{th} \) item with difficulty big for that particular latent class \( g \) correctly given the examinee’s latent ability of \( \theta_{jg} \) in
latent class \( g \). The unconditional probability of a correct response is expressed in the following equation:

\[
p_{ji} = \sum_g \pi_g p_{jig} = \sum_g \pi_g \frac{1}{1 + \exp[-(\theta_{ig} - b_{ig})]}
\]  

(3-8)

where \( p_{ji} \) is the unconditional response probability and \( \pi_g \) is the class mixing proportion; with constraints \( 0 < \pi_g < 1 \) and \( \sum_g \pi_g = 1 \) across classes.

Similarly to the LCA the mixture Rasch model can satisfy the empirical of validating judgment based standard setting procedures by assigning examinees to latent groups which can be compared to the assignments determined by the judgment based methods and other empirical models, yet again the simulated data limited the utility of the mixture Rasch model to providing an achievement level profile for each iteration of simulated data. The resulting class proportions and the corresponding item responses and total scores were collected for each iteration.

**Analysis**

Following the execution of the three models with the 100 simulated data sets the data was compared to identify differences across methods and generate evidence for the use of these empirical models at tools in the validity process of standard setting. The performance of each model was judged in several was. The first metric used to assess whether these models could serve to validate the traditional Expert-based standard setting models was through an evaluation of the average achievement level proportions of each model. Next the standard deviations of the 100 iteration for each model were compared to identify the variance across the methods. An important comparison was made between the average achievement level proportions partitioned
by each model empirical models and the reported achievement level proportions form the assessment that served as a template for the data simulation.

In addition to these base line performance indicators, the distribution of the total scores in relation to their achievement levels was observed to visually represent how the three models partitioned the hypothetical examinees. The distribution identifies visually how the achievement levels are partitioned, but another metric that describes the appropriateness of the clustering solution is Entropy. Entropy is described as the measure of classification certainty (Celeux & Soromenho, 1996). It is a metric that provides evidence for the certainty that the classification assignments were appropriate. Entropy is calculated with the following formula

$$\text{Entropy}(\alpha) = - \sum_{l=1}^{N} \sum_{j=1}^{J} \alpha_{ij} \log \alpha_{ij} \quad (3-9)$$

Relative entropy was used to describe the measure of classification certainty for these models as it is reported by the MPlus software (Muthén and Muthén). Relative entropy is calculated with the formula:

$$E = 1 = \frac{\text{Entropy}(\alpha)}{N \log J} \quad (3-10)$$

An additional performance metric was used to describe the consistency across the methods: The percent agreement across models was calculated. This percent agreement was a simple comparison of whether simulated examinees were assigned to the same achievement levels. The percent agreement across all tree models was calculated as well. Together these values reflect the consistency of achievement level assignment across all the groups.
CHAPTER 4
RESULTS

One of the primary objectives of this study was to incorporate more realistic test data into an evaluation of empirical standard setting procedures. The efforts to achieve realistic conditions were balanced between practical and ideal considerations. The data that was simulated resulted in a 100 hypothetical tests each with 45 item for 180,000 examinees. The simulated data compared well to a realistic testing situation given the practical constraints. The mean, standard deviation, reliability and standard error of measurement for the data that was used as a guide in the simulating of data for this study were 30.42, 8.9, 0.92, and 2.74 respectively. The simulated data resulted in mean, standard deviation, reliability and standard error of measurement values of 28.61, 10.06, 0.92 and 2.84 respectively. The minimum and maximum values for total scores for both the simulated data and the data used as a template were identical at 1 and 45 respectively. Means for Expert-based methods were obtained from (FCAT Yearbook, 2014) and used as population parameters for simulation.

Table 4-1. Average Achievement Level Proportions

<table>
<thead>
<tr>
<th>Univariate MM</th>
<th>Multivariate LCA</th>
<th>Mixture Rasch Model</th>
<th>Expert-based Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL 1 15.111</td>
<td>17.582</td>
<td>17.587</td>
<td>18.690</td>
</tr>
<tr>
<td>AL 3 26.860</td>
<td>23.885</td>
<td>23.885</td>
<td>23.230</td>
</tr>
<tr>
<td>AL 4 26.106</td>
<td>24.244</td>
<td>24.244</td>
<td>23.360</td>
</tr>
<tr>
<td>AL 5 6.528</td>
<td>9.897</td>
<td>9.895</td>
<td>10.210</td>
</tr>
</tbody>
</table>

Note. LCA = latent class analysis, AL = achievement level.

Following considerations regarding the descriptive statistics of the simulated test data the empirical models were compared to meet the second empirical of this study. Performing and evaluating three different models used in previous research provides a platform to evaluate the work in the field thus far. The first performance cue was the
proportion of examinees that were partitioned into achievement levels by the three models.

Table 4-1. shows the proportions of each examinee that were assigned into the 5 achievement levels. The achievement level proportions for the univariate MM ranged from 6.528% for the 5th and highest achievement level to 26.860% for the 3rd achievement level. The range of achievement level proportions for the multivariate latent class analysis was slightly smaller than that of the univariate latent class model spanning from 9.897 for the highest achievement level to 24.392 at the second achievement level. The achievement level proportions for the mixture Rasch model were very similar to that of the multivariate latent class model ranging from 9.895 for the 5th achievement level to 24.388 at the 2nd lowest achievement level. The Expert-based achievement level proportions that were reported in the FCAT 2.0 2013-2014 yearbook ranged from 10.210 at the 5th achievement level to 24.420 at the 2nd achievement level (FCAT Yearbook, 2014). The multivariate latent class analysis and the mixture rash model assigned the simulated students to achievement level proportions that aligned very closely with the reported achievement level proportions. The univariate latent class analysis was not as successful as the mixture rash model and the multivariate latent class analysis, but the class proportions still resembled the proportions reported for the test used as a template for the data simulation.

Table 4-2. depicts the standard deviations of the achievement level assignments for all 100 iterations. The achievement level standard deviations for the univariate MM ranged from 0.0003 for the 5th and highest achievement level to 0.0290% for the 2nd achievement level. The range of achievement level standard deviations for the
multivariate latent class analysis was considerably smaller than that of the univariate latent class model spanning from 0.0009 for the 3rd achievement level to 0.0032 at the 1st achievement level. The achievement level standard deviations for the mixture Rasch model were nearly identical to that of the multivariate latent class model ranging from 0.0009 for the 3rd achievement level to 0.0032 at the 1st achievement level. The Expert-based achievement level standard deviations were not reported that were reported in the FCAT 2.0 2013-(FCAT Yearbook, 2014). The was a noticeable difference in magnitude of standard deviations for the univariate latent class analysis as compared to the multivariate latent class analysis and the mixture Rasch model. This univariate latent class analysis had much larger standard deviations for the first three achievement levels. This suggests higher variation across the 100 iterations. There was a noticeable decrease in variability across the three models. There seemed to be larger variation beginning with the first and second achievement levels with the standard deviations decreasing as the achievement levels increased.

Table 4-2. Standard Deviation of Achievement Level Proportions

<table>
<thead>
<tr>
<th></th>
<th>Univariate MM</th>
<th>Multivariate LCA</th>
<th>Mixture Rasch Model</th>
<th>Expert-based Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL 1</td>
<td>0.0220</td>
<td>0.0032</td>
<td>0.0032</td>
<td>—</td>
</tr>
<tr>
<td>AL 2</td>
<td>0.0290</td>
<td>0.0029</td>
<td>0.0029</td>
<td>—</td>
</tr>
<tr>
<td>AL 3</td>
<td>0.0180</td>
<td>0.0009</td>
<td>0.0009</td>
<td>—</td>
</tr>
<tr>
<td>AL 4</td>
<td>0.0098</td>
<td>0.0010</td>
<td>0.0010</td>
<td>—</td>
</tr>
<tr>
<td>AL 5</td>
<td>0.0003</td>
<td>0.0011</td>
<td>0.0011</td>
<td>—</td>
</tr>
</tbody>
</table>

Note. LCA = latent class analysis, AL = achievement level

Table 4-3. depicts the absolute value of the pairwise differences between the Expert-based achievement level proportions and the average achievement level assignments of the three models employed in this study. The absolute differences in average achievement level proportions between the univariate latent class analysis and
the Expert-based achievement level proportions ranged from 0.975 for the 2\textsuperscript{nd} achievement level to 3.682 for the 5\textsuperscript{th} and highest achievement level. The absolute differences in average achievement level proportions between the multivariate latent class analysis and the Expert-based achievement level proportions ranged from 0.028 for the 2\textsuperscript{nd} achievement level to 1.108 for the 1\textsuperscript{st} and lowest achievement level. The absolute differences in average achievement level proportions between the mixture Rasch model and the Expert-based achievement level proportions ranged from 0.032 for the 2\textsuperscript{nd} achievement level to 1.103 for the 1\textsuperscript{st} and lowest achievement level. Across all three models, the 2\textsuperscript{nd} achievement level proportion had the smallest difference as compared to the achievement level proportions reported in the FCAT 2.0 2013-2014 Yearbook. For the univariate latent class analysis achievement levels 1, 3 and 5 had absolute differences over 3 percent. These differences were the largest at any achievement level across all three models. The differences between the reported Expert-based model proportions and the proportions for the univariate latent class analysis and the mixture Rasch model reveal a much smaller absolute difference than the univariate MM. The largest difference for the multivariate LCA and the mixture Rasch model were 1.108 and 1.103 respectively, all other differences were smaller than 1 percent.

<table>
<thead>
<tr>
<th>Table 4-3. Difference Between Achievement Level Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate MM</td>
</tr>
<tr>
<td>AL 1</td>
</tr>
<tr>
<td>AL 2</td>
</tr>
<tr>
<td>AL 3</td>
</tr>
<tr>
<td>AL 4</td>
</tr>
<tr>
<td>AL 5</td>
</tr>
</tbody>
</table>

Note. LCA = latent class analysis, Al = achievement level
Figure 4-1. shows the distribution of total scores for each achievement level classification under each model. It can be observed that there is little overlap between the achievement level total scores for the univariate MM where as there is noticeable overlap of total scores across achievement levels with the multivariate LCA and the mixture Rasch model. This overlap is the effect of the item level data in the multivariate LCA and the mixture Rasch models.

The measure of classification certainty employed in this study is entropy (Celeux & Soromenho, 1996). It is a metric that provides evidence for the certainty that the classification assignments were appropriate. Entropy values range between 0 and infinity, however, relative entropy indicates classification certainty on a 0 to 1 scale, 1 being absolute certainty and 0 being complete uncertainty (Muthén and Muthén). The relative entropy for the univariate MM model was .999. The entropy for the multivariate LCA model was .999 and the entropy for the picture Rasch Model was also .999.

In addition to performance indices like entropy and the achievement classification proportions, classification accuracy indicated how well the models aligned in their ability to assign the same hypothetical student to the same achievement level classification. Classification accuracy was measured by the percent agreement across each model.

Table 4-4 shows the pair-wise percent agreement for the achievement level assignments of each model. The percentage agreement values represent the proportion of equivalent achievement level assignments between methods. Given that the same simulated examinee response patterns were used in all three models, each examinee was assigned to an achievement level by each model. Their resulting class assignments were compared yielding a fraction of the total number of simulated
examinees that were assigned to the same achievement level. The univariate latent class analysis and the multivariate latent class analysis assigned 85.0% of students to the same achievement levels across all 100 iterations. The univariate latent class analysis and the mixture Rasch model had a percent agreement of 82.0%. The multivariate latent class analysis and the mixture Rasch model assigned 95.0% of the simulated examinees to the same achievement level. When all three models were compared together the overall agreement was 85.3% for all 100 iterations.

Table 4-4. Percentage of Agreement for Achievement Level assignment

<table>
<thead>
<tr>
<th>Method</th>
<th>Univariate MM</th>
<th>Multivariate LCA</th>
<th>Mixture Rasch Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate MM</td>
<td>─</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multivariate LCA</td>
<td>85.0</td>
<td>─</td>
<td></td>
</tr>
<tr>
<td>Mixture Rasch Model</td>
<td>82.0</td>
<td>95.0</td>
<td>─</td>
</tr>
</tbody>
</table>

Note. LCA = latent class analysis
Figure 4-1. Achievement level assignments.
There are several roles that empirical modeling can play in the standard setting procedures and the scope of this study was to address empirical modeling as a source of evidence for validating the current process. This perspectives shapes how the performance of the models was judged and to what extent the results met their intended function.

Data Simulation

An evaluation of this study begins with the simulated data used to mimic a real high-stakes testing situation where standards were set partitioning examinees into achievement levels. Given the current standard setting procedures no validity evidence for that process involves empirically partitioning the examinees scores to identify clusters of student achievement. Advancing the field of study surrounding using statistical models that cluster objects or individuals into the standard setting process involves experimenting with these models in realistic testing environments. Although real data were not available, the use of real test characteristics and item parameters in the simulation process enabled a practical attempt at exploring how the empirical models can participate in the standard setting process for large-scale high-stakes assessments.

To address the utility of a clustering model for obtaining validity evidence it is important to observe the models in the setting of its desired use. Principles of validity developed by Cronbach (1955) and Messick (1980) instruct that validity is a characteristic of scores and outcomes rather than of tests. In order to provide sufficient evidence for the validity, each situation in which an assessment is given must be
validated appropriately. The field of study surrounding empirical modeling in the standard setting process had yet to employ these models in a high-stakes large-scale environment. This study begins that effort by using a real high-stakes assessment as a template for simulating data.

The mean total scores of the simulated tests, 28.61, fell within a realistic range of test scores observed in the detailed reports used as the template for simulation (FCAT Yearbook, 2014). The standard deviation of the simulated tests, 10.06, was also within a realistic range depicted in the report from the state of Florida (FCAT Yearbook, 2014). The reliability and standard error of measurement for the simulated tests were nearly identical to the reported values in the detailed report from the state of Florida. These descriptive statistics from the test simulations support that this data represents a test like situation that is similar to the real high-stakes large-scale testing assessment used by the state of Florida.

Using a real high-stakes, large-scale assessment as a template for data simulation built upon the previous research in this field of study. Brown (2007) utilized a small sample of items from a large-scale assessment, but this assessment was non-mandatory (Brown, 2007). This gave that particular study real student scores to use in the LCA. The real student scores were also subjected to the traditional standard setting procedures providing a unique experimental design. An empirical method could directly be compared to the traditional standard setting procedure, however collecting actual scores limited the number of students in the sample and limited the number of items. This study tested only 181 students across 12 items. The number of students used in this study poses the issues of both methodological appropriateness and practical
relevance. Nylund (2007) suggests that the performance of a LCA is effected by the sample size with specific regard to categorical data. The number of subjects in Brown’s (2007) study could compromise the accuracy of the LCA. It has been show that the ability to select the correct number of classes in simulation was only 1% for a model with \( n = 200 \) and 56% for \( n = 500 \) (Nylund, 2007), while, nearly 100% for \( n = 1000 \).

Additionally, the comparison of the sample of items and students used by Brown (2007) pales in comparison to the magnitude of nearly all state assessments. The assessment used as a template in this model reach really 200,000 students at each grade level (FCAT Yearbook, 2014). The data simulated in this study aligns with the logic of Brown (2007) to use empirical modeling with a large-scale assessment, but observes model performance in a simulated environment that more closely resembles a large-scale assessment as well as better meeting the needs of the empirical models described by Nylund (2007).

Similarly to this study, Jiao (2011) utilized simulated data for their evaluation of empirical modeling in the standard setting process. They simulated data for 10,000 hypothetical students, with an ability level distribution with 5 achievement groups with unique mean and standard deviations. The number of simulated students exceeds the number of students used in the study by Brown (2007) and Sireci (1999). This is larger sample is a step in the direction of realistic large-scale high-stakes assessments, but there were no specific reference to where the parameters were obtained. An additional similarly between the data from Jiao (2011) to the data used in this study was the ability level distribution with 5 distinct levels used to generate student level item responses. However, the lack of any theoretical or practical reasons for determining the distribution
used leaves lingering questions about the environment to which this simulation represents. Although all situations are limited, the data generated in Jiao’s (2011) appears arbitrary without any relevant comparison limiting the scope of that study even further. Revealed by the author was the need to incorporate realistic data for large-scale high-stakes assessments (Jiao, 2011). This study addresses that call and addresses explicitly where the simulated parameters were obtained. Simulating a larger sample than that used in Jiao (2011) reflecting a real testing environment demonstrates the emphasis on creating a simulated environment that closely mimics that of a real large-scale high-stakes assessment that has yet to be addressed in any previous work (Jiao, 2011; Brown, 2007; Sireci, 1999).

Sireci (1999) was the first to introduce the concept of empirical modeling into the standard setting procedures prior to critical changes in the educational system in the US. Given the time period of this study, the context of the standardized assessment may have been different, but nonetheless the logic from that study was incorporated into one of the models used in this study. The decision by Sireci (1999) to incorporate total scores into a clustering models was used by one of the models in this study, but considering the shifts in the national educational system primarily attributed to the NCLB Act in 2001, the data used in this study was different than what was used by Sireci (1999). The data used by Sireci (1999) was real student scores from a state-wide assessment used to identify remedial students in mathematics (Sireci, 1999). This assessment was over 100 items in length and used to partition students in to three achievement levels. Although this differs considerable from the data used in this study, the logic for the model based on this study was to use total scores as the clustering
variable. In order to use total scores to categorize students into achievement levels and compare those results across methods, the same data must have been consistent across models. Therefore the item responses generated in the simulation were summed to create a single total score to be used in the univariate latent class model. These data was simulated to resemble an assessment that exists in the current educational system rather than one from the pre millennial educational system like the environment existing at the time of Sireci's (1999) study.

The item parameters and test characteristics that were used were from a one example of a large-scale high-stakes testing environment. This information was limited in several ways. The criteria used in the simulation were limited first by what was publicly reported by the state of Florida. These reports published by the Florida Department of Education (DOE) are detailed, but not comprehensive constraining the simulation process. Given that this study employed empirical clustering procedures in a simulated test environment, the underlying groups of students was determined a priori. To produce a simulated environment that exactly reflected the state of Florida assessments, the mean and standard deviations of each achievement level would be needed. The logic behind separating examinees into discrete groupings based on the underlying ability level means that five groups with unique distributions exist within the sample itself. Knowing the mean and standard deviation of ability for each of the groups would allow for a precise replication of the template exam (simulating mixture IRT data). However, this information was not provided in the detailed reports, and therefore the means of each achievement level had to be estimated by the total ability level distribution that was given in the report.
An additional limitation in regards to the specific item and test characteristics relates to how those characteristics reflect just one of many high-stakes large-scale testing environments. More likely than not each state educational system has unique assessments. This study is the first to explore using empirical models for standard setting with large-scale high-stakes assessments but it is limited to a simulation that reflects the reading assessments used in the state of Florida. The assessment that was used as a template was from the FCAT 2.0 program that was replaced following the 2013-2014 school year (Changing). The new assessment is different from the FCAT 2.0, so the degree to which the results from this simulation relate to that test are unclear, but beyond limiting the inferences that can be made from this study, it shows the dynamic nature of high-stakes testing and the accountability system in general. It would have been far too overwhelming to attempt to study all the different forms and types of assessment within one state let alone all the educational systems across the US.

**Model Comparisons**

Accepting the limitations in reach and taking this study for what it is, allows for the performance of the models to be judged and meaningful conclusions to be drawn. The second empirical of this study was to compare the performance of three models using the same data to provide a frame of reference for the field indicating whether there were any noticeable differences in performance from the models that have been utilized in previous studies. Model performance was judged on how well the resulting achievement level assignments reflected the proportions from the real testing situation that was used as a template for simulating the data. Judging the achievement level proportions from the empirical models and the real testing situation emulates how the
models could be used as validity evidence in a true life application. Each empirical model was compared to the reported achievement level proportions by the Florida DOE. Additionally model consistency between the empirical models was observed to address the discrepancies between the models themselves.

The averaged achievement level proportions from all 100 iterations of the simulation were very close to that of the achievement level proportions reported by the Florida DOE. The univariate MM partitioned the simulated examinees into achievement levels that were the furthest from the reported achievement level proportions. The multivariate LCA and the mixture Rasch Model were equally close to the reported achievement level proportions. Table 4-3 depicts the pair-wise differences between the achievement level proportions of the empirical models and the achievement level proportions from the Florida DOE reports.

The values in Table 4-3. show how close the achievement classifications aligned with the achievement level proportions from the Florida DOE reports. Given the closeness in achievement level proportions it can be inferred that these three methods were all successful in partitioning the simulated examinees into groups that closely reflected the template. This would suggest that these models can play a role in validating the standard setting process. This study has implications that are limited to this specific circumstance, meaning that this study confirms that these models could play a role in providing validity evidence for a similar test and educational system. Carrying inferences beyond this specific situation would not be supported by this study, however, the success of his study suggests considerable promise for the use of these
models in situations beyond the one in this study. Further research would extend this study to other types of assessments in other educational systems.

In addition to evaluating the differences in the proportions of class assignments, the relative consistency of each model can be a means to describe the performance of the empirical models. Consistency of models represents the degree to which the models assign the same simulated examinee to the same achievement level. The absence in the literature of studies comparing the performance of one empirical model to another leaves no means to address which model if any perform better in the context of providing validity evidence. An important measure of performance was the percentage of examinees assigned to each achievement level. However, similar percentages of examinees in each achievement level do not confirm that the same hypothetical examinees were assigned to the same achievement levels across groups.

To add more support for the models as tools to validate the current standard setting procedures, having models that assign the same individuals to the same achievement level demonstrates their ability to serve as validity tools. We can confirm that if the models provide consistent class proportions and those classes are made up of the same individuals that those methods perform consistently. Table 4-4. shows the pair-wise percent agreement between empirical models. The univariate MM and the multivariate LCA had an 85% agreement. This means that 85% of the students with the same item responses were assigned to the same classes with these two models. The univariate MM and the Mixture Rasch Model had an agreement of 82% meaning that 82% of the students who earning the same test scores were assigned to the same classes with these two models. The multivariate LCA and the Mixture Rasch Model had
the highest agreement of 95%. Only 5% of the simulated examinees with the same item responses were assigned to different achievement levels. These values are relatively high, but comparisons to the percent agreement for the Expert-based methods a reported by the Florida DOE shows that these percent agreement values are substantially larger (FCAT Yearbook, 2014). The state of Florida describes the accuracy of classification as a metric developed by Livingston and Lewis (1995), which defines accuracy of classification as the extent to which actual classification agrees with the classification that would be made if the true classification could be known. Following the basis of test and measurement theory, a test is a representation of a student true ability, but given that there is not perfect test there always exists some level of error in that measurement. The method used to calculate the accuracy of classification makes a statistical estimation of the true classifications that are represented in each classifications of achievement based on observed examinee scores. Accuracy of classification represents the percent of students in each observed achievement level that coincide with the true classification (Livingston and Lewis, 1995). Table 5-1 shows the percent agreement for the eighth grade levels who administered the Reading FCAT 2.0 assessment in the 2013-2014 school year (FCAT Yearbook, 2014).

<table>
<thead>
<tr>
<th>Grade</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.716</td>
</tr>
<tr>
<td>4</td>
<td>0.692</td>
</tr>
<tr>
<td>5</td>
<td>0.686</td>
</tr>
<tr>
<td>6</td>
<td>0.701</td>
</tr>
<tr>
<td>7</td>
<td>0.693</td>
</tr>
<tr>
<td>8</td>
<td>0.671</td>
</tr>
<tr>
<td>9</td>
<td>0.680</td>
</tr>
<tr>
<td>10</td>
<td>0.667</td>
</tr>
</tbody>
</table>
The consistency values of the empirical models used in this study all have accuracy percentages that are much higher than the accuracy reported by the state of Florida. The accuracy percentages reported by the Florida DOE are quite low, supporting the need for better standard setting models. The empirical models operate with much higher accuracy and although this study cannot confirm that these models would perform better in situations outside the scope of this study, the success of these models in this situations calls for further investigation and a dedicated effort to incorporate empirical modeling into the standard setting process.

These average class proportions, pairwise differences between model including the Expert-based achievement level proportions reported by the Florida DOE and the percent agreements of the empirical models used in this study can be compared to similar results from the previous studies that employed the same models. This analysis can illustrate how well the results of this study aligned with the results from those previous studies, as well as, provide evidence for which models performed better.

Sireci (1999) used total scores to cluster students and compared the resulting cluster proportions to the actual achievement level proportions for the assessment used obtained locally. Similarly to this study, the only data that could be used to compare the empirical methods with the traditional standard setting procedures were the achievement level proportions. Sireci (1999) used two slightly different clustering techniques revealing two separate achievement level profiles for three achievement levels with proportions of 32.0, 33.5, 34.6 and 31.0, 31.2, 37.8 respectively. The achievement level proportions from the traditional Expert-based standard setting procedures were 27.2, 40.2, 32.6 (Sireci, 1999). The achievement level proportions from
the cluster analysis aligned well with the actual achievement level proportions. Similar results were observed with the univariate latent class analysis in this study. The success of the univariate cluster analysis with total scores from Sireci (1999) was replicated by the results of this study suggesting that using total scores to cluster examinees can be successful in a small scale environment, as well as, a large scale environment. This is promising for the use of empirical models in the standard setting process as the simplest clustering model proved to be relatively successful.

The study by Sireci (1999) was similar to the univariate latent class analysis used in this study which allows for the comparisons of results. This comparison provides support for the results of both studies, in addition to, the use of univariate cluster models in this field of study. Comparing the results of the multivariate latent class model and the results from the latent class model used in Brown (2007) further substantiates the previous work in the field and demonstrates the effect of item level data on the model performance.

Brown (2007) was a uniquely designed study in that a specific selection of items were given to a sample of students generating a unique sample of item responses without an a priori designation of achievement level proportions. This gave Brown (2007) more approaches to assess model performance, however, ultimately achievement level proportions could be compared between the traditional Expert-based standard setting methods and the empirical LCA. Brown (2007) reported percent agreement across the methods used in that study. The traditional Expert-based standard setting procedures produced equivalent achievement level proportions as compared to the LCA model for 92.2% of the students in the study. Although there is no
direct comparison to this study given that the only data that was available for the real
student achievement levels were total level proportions, the high level of accuracy seen
in Brown’s (2007) study and the closeness of the achievement level proportions
partitioned by the multivariate LCA and the reported proportions by the Florida DOE in
this study suggest that the multivariate LCA in both model showed constant results. The
ability of the multivariate LCA models to partition data according to ability level seems
promising given the results of both Brown (2007) and this study. Having Brown (2007)
collect real data and execute a traditional standard setting procedure on the same data
demonstrates the utility of the multivariate LCA in the standard setting process from a
real applied situation. This study explores the multivariate LCA in a different way, but
demonstrates that the model can achieve similar results in a large-scale setting which
was a primary empirical of the study. Together these two studies suggest that the
multivariate LCA is a rich platform for continued exploration for the field.

The empirical models used in this study are designed to detect differences
between groups. The literature has suggested that these models function most
effectively when the separation between the groups is larger (Sireci, 1999). Fortunately,
the test that was used as a template had a wide distribution of scores enabling for
groups that were more distinctly separated. The entropy values from the simulated data
of .999 are very high. This shows that the achievement level classifications were easily
distinguished. This may not be true for all high-stakes assessments. It is important to
hypothesize the separation between groups when applying these empirical models to a
testing situation. An assessment in which the examinee abilities are tightly clustered
could generate different results than the results from this study. The performance of the
empirical models relating to the purpose of this study in a situation where the construct of interest is measured with items that are consistently easy or consistently challenging for the sample may result in scores that are tightly clustered affecting the ability to make appropriate distinctions between the ability levels. However, the Expert-based methods employed in the current standard setting process could more appropriately adapt their partitioning methods to account for a tightly clustered group. This is a limitation of using empirical models. Improvements in cluster modeling techniques could provide more useful tools in situations where individuals within a sample are tightly clustered.

The more recent study to date in the field of empirical standard setting procedures is the simulation study by Jiao (2011). This study found that a mixture Rasch Model can effectively retrieve the class proportions designated a priori in simulation (Jiao, 2011). The true achievement level proportions used in the simulation were 6.00%, 13.00%, 53.00%, 22.00% and 6.00% for the 1st through 5th achievement levels respectively (Jiao, 2011). The mixture Rasch model partitioned the simulated data into achievement levels with proportions 6.02%, 12.16%, 53.45%, 22.44% and 5.93% for the 1st through 5th achievement levels respectively (Jiao, 2011). Although, the proportions are naturally different the performance of the mixture Rasch model was very similar to that not the same model in this study. They achievement level proportions of the mixture Rasch model were very similar to the proportions of the reported achievement levels used to simulate the data. The most significant addition to the field that this study makes in regards to using a mixture Rash model is the explicit comparison to a real assessment setting. As mentioned in the discussion about the data accumulation, Jiao (2011) made no direct reference to the parameters used to simulate the data. That
study was important in employing a new method of partitioning students and made reference to using real assessment like parameters, but having not direct comparison leaves the results seeming arbitrary. This study confirms the potential for such a mixture Rasch model as a tool for validating current standard setting methods, but utilizing real assessment conditions in the simulation. Nonetheless, the consistence in performance suggests the continued use of the mixture Rasch model in the standard setting process.

It is important to note that the mixture Rasch model estimates class proportions using a single parameter, the difficulty parameter. The data used in this study was generated with three item parameters discrimination, difficulty and guessing. The simulated data reflects characteristics of all three parameters. However, the mixture Rasch model using just one parameter, was sufficient in assigning achievement level proportions that were similar to proportions of the template data and the proportions derived by the other models.

An addition element of this study that different from the traditional standard setting process is the use of scaled scores to separate students in the current standard setting process. Total scores are converted to a scaled score and the achievement level cut offs, as determined by the standard setting procedure, are also converted to a scaled score (FCAT Yearbook, 2014). This means that achievement level distinctions are made using these scaled scores. For this study, raw scores were used to separate individuals. This is not theoretically limiting for this study and does not serve to change the implications of the results, however, it limits the comparisons that can be made. The cut offs between the different models and the real cut offs used in the realistic situation cannot be compared because one is using raw scores and the other is using a scaled
score. It is important to address this in future research to show that the cut off distinctions observed by the empirical models and the current standard setting procedures coincide. This once again shows a limitation of the reported data and test.
CHAPTER 6
CONCLUSION

The importance of the standardized assessment in the US educational system is undeniable. The decisions that are made across the educational spectrum as results of standardized assessments demonstrate this importance (Sciarra, 2015). But along with this significance comes controversy and resistance (Oliver, 2015). The polarizing issues with standardized testing are the consideration for the uses of standardized test scores (Baker, 2010; Turniseed, 2015). These considerations rely heavily on the ability to determine achievement levels among students. The criticism and lack of improvement that has plagued our educational system for some time directs our judgment to nearly all aspects of the system, but one of the facets most appropriately scrutinized, is the process for determining the achievement levels.

There is considerable evidence to show that the standard setting process has limitations. The self-reported errors in classification are one, the lack of empirical validity metrics another. Although these errors exists, so too are there statistical tools that function to detect, minimize and supplement the standard setting process. Identifying a problem, isolating the cause and providing potential solutions should be a regular occurrence in our educational system. A field of study has emerged with those attributed for the standard setting process. Using clustering models to partition students into achievement level groupings is an attainable task. This study and the several that have come before demonstrate the possibility of empirically partitioning students into such groupings. Tools exist to supplement the current standard setting process and the effects of their implementation can be significant.
An important element of making decisions is the assumption that the information used to make such decision was accurate and meaningful. Accepting the decisions that are made across the educational spectrum makes the assumption that the achievement level distinctions that result from large-scale high-stakes assessments are valid and meaningful. With the alarmingly low levels of accuracy in classification and the errors that exist in all expert-based based classifications that dominate the current standard setting procedures, the validity and meaning of the achievement level classifications is far from solidified. Granted these methods have face-validity and are rigorous, the most practical utilization of the empirical-based methods would be to incorporate them at the various stages of the standard setting process. Using the empirical-based methods during the traditional standard setting procedures offers increased objectivity to the process and allows for there to be an iterative cooperation between the two models. The expert-based practices already contain stages of standard setting development which could allow for the incorporation of the empirical models at the different stages.

The utilization of the empirical models throughout the existing standard setting process could streamline the current process. The use of the empirical models has limitations and the field has yet to process to a point where these models could appropriately serve as an alternative practice, however, their performance as demonstrated by this study and others show their functionality in determining achievement level distinctions. However, appropriate inclusion into the current models is needed to account for the limitations, as well as, the status quo of standard setting. Nonetheless, addressing the importance of the standard setting process and
progressing the field of empirical standard setting is a worthwhile cause for nearly all stakeholders of the educational system.
LIST OF REFERENCES


Fla. Stat. § 1008.22

Fla. Stat. § 1008.34

Fla. Stat. § 1012.34

Florida Statewide Assessments 2014 Yearbook


BIOGRAPHICAL SKETCH

Cameron Bernard Martin was born outside of Buffalo, NY. His family lived there for a short time before relocating to Clearwater, FL. His family remained there until his early adulthood.

He was enrolled at Guardian Angels Catholic School from kindergarten to eighth grade and then studied at Clearwater Central Catholic High School before his acceptance at the University of Florida. At the University of Florida, Cameron earned a bachelor’s degree in zoology.

His future is carefully planned, yet appropriately dynamic. He hopes to make a significant impact in the world and strives to achieve greatness in all that he pursues.