Title

An Application of a Structured Mixture Rasch Model to Computer Adaptive Data: An Example in Early Kindergarten Geometry

Author

Langi, Meredith

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

An Application of a Structured Mixture Rasch Model to Computer Adaptive Data: An Example in Early Kindergarten Geometry

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Meredith Lindsay Langi

2021

ABSTRACT OF THE THESIS

An Application of a Structured Mixture Rasch Model to Computer Adaptive Data: An Example in Early Kindergarten Geometry

by

Meredith Lindsay Langi
Master of Science in Statistics
University of California, Los Angeles, 2021
Professor Chad Hazlett, Chair

In the field of education, understanding differences in student performance by content area subdomains, and classifying students based on these differences, is of interest for differentiating instruction. Structured mixture item response theory (IRT) models offer a unique opportunity to achieve these goals within a confirmatory modeling approach. However, to date, there have been no known applications of this type of model to computer adaptive testing (CAT), a common test design in large-scale educational assessments. This thesis fills this gap by demonstrating the application of a particular structured mixture IRT model to early kindergarten geometry data. Results suggest the model is useful in CAT applications for understanding how students differ by domain, but that the test design must follow certain specifications for student classifications.

The thesis of Meredith Lindsay Langi is approved.

Minjeong Jeon

Ying Nian Wu

Chad Hazlett, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Item Response Theory (IRT) is a modeling framework that is widely used in educational contexts for estimating student ability on important academic and non-academic constructs. Under this framework, IRT models explain observed responses on an assessment as a function of student ability and assessment item properties. Most IRT models assume that students are from a homogeneous population, but this assumption may be violated either through known or unknown clustering (Jeon, 2018). For example, due to a variety of circumstances, students often enter kindergarten with different levels of mathematics achievement (Kuhfeld, Soland, Pitts, & Burchinal, 2020; Reardon & Portilla, 2016; von Hippel, Workman, & Downey, 2018; Wolf, Magnuson, & Kimbro, 2017). This could be due to different opportunities and varying exposure to different areas of mathematics, making some domains of mathematics more difficult for some groups student. When group membership is unknown and inferred from the data, mixture IRT can be used to model potential heterogeneity. Specifically, mixture IRT models assume that students are drawn from multiple latent populations and these differences impact the relationship between the items and the observed responses. Several applications of mixture IRT models exist in the education literature (e.g., Bolt, Cohen, & Wollack, 2002; Mislevy & Verhelst, 1990), but this paper focuses on a mixture IRT model that uses item information to differentiate between latent classes. This model was originally termed the Saltus model (Wilson, 1989) and has been extended and applied across a variety of contexts (e.g., Jeon, 2018; Jeon, Draney, & Wilson, 2015; Mislevy & Wilson, 1996). A key feature of this model is that the item information allows for the exploration of heterogeneity in students by important curricular areas. This makes it very useful in exploring differences among students

based on instructional areas, and thereby providing curriculum relevant feedback that can be used to support instruction. To date, there are no known applications of this structured mixture IRT model to computer adaptive data. This thesis demonstrates the application of this model to CAT data by exploring differences in early kindergarten geometry ability based on a widely used education assessment, the MAP Growth K-2 mathematics assessment.

This thesis is organized as follows. In this first chapter, I briefly discuss the importance of early kindergarten geometry skills, as well as introduce the details of the MAP Growth K-2 math assessment. In the second chapter, I briefly review other psychometric models that have been, or could be, applied in similar contexts. In the third chapter, I review the methods used in the analysis. In the fourth chapter, I present results. Finally, I present a discussion of the implications in the fifth chapter.

1.1 Early Kindergarten Geometry

Math skills in a variety of domains at early childhood have been shown to be an important predictor of future mathematics performance and are essential foundational skills that are necessary for math development (e.g., Newcombe & Frick, 2010; Verdine, Irwin, Golinkoff, & Hirsh-Pasek, 2014; Watts, Duncan, Siegler, & Davis-Kean, 2014). As mentioned, students enter kindergarten with different levels of mathematics exposure and knowledge, which results in differing instructional needs for students. Since teachers' views of children's math skills are not always aligned with student performance (Abry, Latham, Bassok, & LoCasale-Crouch, 2015), assessments should provide instructionally actionable feedback regarding individual student's knowledge on key instructional areas. The structured mixture IRT model is a useful tool in supporting differentiated instruction because it can link items to instructional areas and provide insight into domains that differ most among early kindergarteners.

This thesis focuses on student performance in geometry and its subdomains at

kindergarten entry. Specifically, this paper uses the classification of geometry subdomains by the Common Core State Standards (CCSS). CCSS are a national set of standards that states can choose to adopt and adjust for their own contexts. Most US states (41), territories (4), and the District of Columbia have adopted these standards, meaning that the majority of educators in the United States use them to guide instruction. In CCSS, kindergarten geometry has two subdomains: (1) identifying and describing shapes (referred to as identifying shapes for brevity), and (2) analyzing, comparing, creating, and composing shapes (referred to as analyzing shapes). Table 1.1 shows these subdomains and the associated standards. Since there is a CCSS-aligned MAP Growth K-2 assessment that contains item linkages with these standards, we can make use of this information by applying the structured mixture IRT model to understand how students may differ on these two subdomains.

1.2 MAP Growth K-2 Assessment

This study uses data from NWEA's MAP Growth K-2 interim math assessment, an assessment that is administered on a computer or tablet to students in kindergarten to second grade. The assessment typically takes students less than 30 minutes and provides audio support since reading skills vary widely at this age. MAP Growth utilizes a "cross-grade vertical scale that assesses achievement according to standards-aligned content" (NWEA, 2019, p.9). The assessment is typically administered 3-4 times per year to allow for tracking student growth, although only one test occasion is used for this study. Three key components of the assessment design are important to review for this analysis: (1) the test blueprint, (2) the psychometric model, and (3) the adaptive approach. While some of the details discussed in this section are specific to this particular assessment, the overarching issues are typical in large CAT assessments and would apply to other applications of this structured mixture IRT model.

Table 1.1: CCSS Kindergarten geometry subdomains and associated standards

Subdomain	Standard
Identify and Describe Shapes	CCSS.MATH.CONTENT.K.G.A.1: Describe objects in the environment using names of shapes, and describe the relative positions of these objects using terms such as above, below, beside, in front of, behind, and next to. CCSS.MATH.CONTENT.K.G.A.2: Correctly name shapes regardless of their orientations or overall size. CCSS.MATH.CONTENT.K.G.A.3: Identify shapes as two-dimensional (lying in a plane, ”flat”) or three-dimensional (”solid”).
Analyze, compare, create, and compose shapes	CCSS.MATH.CONTENT.K.G.B.4: Analyze and compare two- and three-dimensional shapes, in different sizes and orientations, using informal language to describe their similarities, differences, parts (e.g., number of sides and vertices/”corners”) and other attributes (e.g., having sides of equal length). CCSS.MATH.CONTENT.K.G.B.5: Model shapes in the world by building shapes from components (e.g., sticks and clay balls) and drawing shapes. CCSS.MATH.CONTENT.K.G.B.6: Compose simple shapes to form larger shapes. For example, ”Can you join these two triangles with full sides touching to make a rectangle?”

1.2.1 Test Blueprint

A test blueprint is a guide to determine the structure and content of the assessment. This blueprint is used to determine the appropriate number of items within each domain to administer to the student during the test occasion. In the CCSS-aligned MAP Growth K-2 math assessment, there are four domains: (a) Operations and Algebraic Thinking, (b) Numbers and Operations, (c) Measurement and Data, and (d) Geometry. The blueprint specifies the number of items per domain that each student will see. However, each domain is further divided into subdomains, and the test blueprint does not specify the number of items per subdomain. In other words, there is a set number of geometry items that students are required to take, but there is no determination as to how many of these items will be of each subdomain. Additionally, items are carefully linked to the common core standards by content specialists, but these items were not specifically designed to measure each standard or CCSS subdomain. The implications of this design will be discussed further in Chapter 3.

1.2.2 Psychometric Model

The MAP Growth K-2 assessment is designed using a specific IRT model, the Rasch model (Rasch, 1960). In the Rasch model, the probability of a correct response on item i for student j is given as,

$$P(Y_{ij} = 1|\theta_j) = \frac{exp(\theta_j - \beta_i)}{1 + exp(\theta_j - \beta_i)}, \tag{1.1}$$

where θ_j is the latent math ability for student j and β_i is the item difficulty for item i and is equal to the point on the ability scale at which a student would have a 50% probability of responding correctly. Importantly, it is assumed that θ_j is a continuous, normal distribution with a single mean and variance ($\theta_j \sim N(\mu, \sigma^2)$). The benefit of using this model is that the latent ability and item difficulty are on the same scale. The difference between two student ability scores is the same regardless of the difficulty of

items that the student is administered. The difference between two item difficulties is also the same throughout the entire scale (NWEA, 2019).

The scale of θ_j (latent ability) and β_i (item difficulty) is on the logit metric, which is not necessarily interpretable in common educational settings. Therefore, a linear transformation is used where,

$$RIT_j = (\theta_j \times 10) + 200. \tag{1.2}$$

The RIT (Rasch unIT) score is used on all MAP Growth tests and ranges from 100 to 350 (NWEA, 2019). In this paper, both scales will be used in the presentation of results.

1.2.3 Computer Adaptive Design

Computer adaptive testing is an assessment approach that allows for shorter tests and more accurate estimates of student ability (Linden, van der Linden, & Glas, 2000). The basic idea is that a student's correct response leads to a more difficult follow-up question, whereas an incorrect response leads to an easier follow-up question. Two key components in an adaptive assessment are the ability estimate and the item selection procedure. In this assessment, the ability estimate is updated after each item response following an approximate Bayes procedure proposed by Owen (1975). For each estimate, all items up to that point are used in the estimate (NWEA, 2019). Once the ability estimate has been updated, a new item is selected from the item pool that is the appropriate level and fits within the blueprint design. The test ends when the ability estimate meets a pre-established threshold for measurement precision and the number of items satisfies the blueprint specifications. For a more detailed discussion of CAT assessments and the technical components of NWEA's assessment, see Van der Linden and Pashley (2009) and NWEA (2019), respectively.

As a result of the item selection process, students are administered some items but not others. This leads to a rather large and sparse data matrix that presents challenges

for future item analysis. The adaptive nature also means that students at differing levels of ability are administered items with differing levels of difficulty.

1.3 Chapter Conclusion

Strong kindergarten geometry skills provide an important base in developing future math skills. Understanding how students differ on these skills is important for differentiating instruction, and tying these differences among students to specific instructional standards allows teachers to make the necessary adjustments. Any modeling process for understanding differences based on subdomains of geometry must take into account the original assessment design. In the case of computer adaptive tests, it is important to consider the test blueprint, the psychometric model used for test development (which is most often an IRT-based model), and the adaptive nature of the assessment. The next chapter reviews possible psychometric models and their strengths and limitations, and emphasizes how the specific structured mixture IRT model is well suited for this particular context.

CHAPTER 2

Review of Psychometric Classification Models

In the field of psychometrics, there are several possible models that can be used to explore differences in student performance on subdomains of kindergarten geometry. Each of these models (or groups of models) offers different interpretations to the question of differences in student performance. However, due to the adaptive design of the MAP Growth assessment, many of these models are not appropriate in this application. This section briefly reviews some common groups of psychometric models and discusses their strengths and limitations in the context of this research. These strengths and limitations are summarized in Table 2.1.

2.1 Diagnostic Classification Models

Diagnostic Classification Models (DCMs, often called Cognitive Diagnostic Models or CDMs) are a type of latent class model. Latent class models classify students into classes based on students' responses to test items. In other words, the latent classes are unobservable and are inferred from the data. DCMs impose specific confirmatory constraints so that students can be assigned to latent classes based on (typically) mastery or non-mastery of important subdomains (Rupp, Templin, & Henson, 2010). For example, were a DCM to be applied to this context, we would learn which students have mastered the relevant CCSS domains, and which students need more targeted instruction in certain domains. This is clearly useful information for teachers to have at the time of Kindergarten entry, however, several limitations prevent DCMs from being applicable in this way in the context of MAP Growth. First, DCMs function best when test items are specifically

Table 2.1: Summary table of strengths and limitations of model groups for application in current analysis

Model Type	Classifies Students	Models Sub-domains	Maintains key assumptions of MAP Growth	Manageable with large data	Used in CAT applications
Latent Class Models	Yes	Yes		Yes	Yes
Explanatory IRT Models		Yes	Yes		Yes
Exploratory Mixture	Yes		Yes		Yes
Confirmatory Mixture IRT	Yes	Yes	Yes	Yes	

designed to test the domains, referred to as attributes, of interest (Rupp & Templin, 2008). In the MAP Growth assessment, a continuous, unidimensional latent math construct is assumed, whereas DCMs are multidimensional in nature. Relatedly, the information linking the items to the CCSS was developed retroactively, meaning the items may not clearly measure a specific domain or attribute. Second, DCMs are built on the assumption that latent classes are categorical and that there is no variation within the classes. This is very different from the assumption of a continuous latent trait that underlies the development of MAP Growth and many other CAT assessments. Taken together, these limitations of DCMs in this context would lead to unsatisfactory and inaccurate student classifications (Gierl, 2007; Gierl, Leighton, & Hunka, 2007). While some authors argue that there may be contexts where retrofitting is acceptable (e.g., Liu, Huggins-Manley, & Bulut, 2018), DCMs are not considered an ideal approach for student classifications in this context.

2.2 Explanatory IRT Models

Explanatory IRT (De Boeck & Wilson, 2014) is a useful approach to understanding how performance differs by subdomains or other item characteristics. While these models do not classify students based on performance, they are still included in this review as they model how differences in probability for a correct response to an item based on subdomains. (Note that Wilson and De Boeck also discuss person explanatory models, but those differ from the goals of this analysis and are not included here.) One specific example of an item explanatory model that is relevant for this paper is the Linear Logistic Test Model (LLTM; Fischer, 1973). This model is similar to the Rasch model in Equation 1.1, except the item difficulty parameter, β_i, is considered a linear combination of the contributions of various item properties. This model was an important step in developing psychometric models that connect the theoretical cognitive constructions (item characteristics) with the empirical measurements (Embretson & Gorin, 2001). However, the contribution of specific item characteristics to the item difficulty does not vary by

person or latent classes, and as such, there is no classification of students based on performance (Hartz, 2002; Stout, 2007; von Davier, Xu, & Carstensen, 2009). This means that the results from an analysis using LLTM, as well as other item explanatory models, are less useful for teachers interested in directly targeting instruction based on differences in performance on subdomains or item characteristics.

2.3 Exploratory Mixture IRT

Broadly, mixture IRT combines classification models with item response theory by incorporating latent classes into the item response theory measurement model allowing the IRT model to vary across classes (Gnaldi, Bacci, & Bartolucci, 2016; Rost, 1990; Smit, Kelderman, Flier, et al., 2000). Mixture IRT is most often applied using an exploratory approach. In other words, the number of latent classes is unknown and the goal of the analysis is to determine the number of latent classes represented in the data. The flexibility of mixture IRT allows for differentiating among students not only at different performance levels, but also across different response patterns. However, the exploratory nature of the latent classes makes the substantive interpretation of the classes difficult for practical applications. Additionally, mixture IRT with large samples is extremely difficult to estimate, making it difficult to apply in the current study.

2.4 Confirmatory Mixture IRT

The difference between confirmatory mixture IRT and exploratory mixture IRT is that the number of latent classes is defined a priori based on theoretical or practical considerations. Confirmatory mixture IRT is less common than the exploratory approach but a few examples do exist. For example, Bolt, Cohen, and Wollack (2001) use mixture Rasch modeling to explore test speededness in respondents and Mislevy and Verhelst (1990) use mixture IRT to identify solution strategies employ by test takers. The benefit of confirmatory mixture IRT is that it allows for the flexibility of the measurement model

across latent classes, while reducing the estimation burden by defining the number of latent classes based on theoretical or practical considerations. The Saltus model is an example of a confirmatory mixture IRT model. It was initially developed in order to understand cognitive leaps in development and to assign students to Piagetian type stages, although its use can be extended beyond this original purpose (Jeon, 2018; Jeon et al., 2015). It incorporates latent classes based on an a priori defined theory, making the interpretation of student classifications useful for instructional interventions, as well as making it easier to estimate. The Saltus model also incorporates item characteristics, similar to the LLTM, such that the definition of the latent classes is tied to differences in performance based on the specific item characteristics. The incorporation of the item characteristics imposes a specific structure on the model, making it confirmatory in the sense of its item structure. The Saltus model is an extension of the Rasch model, meaning that it maintains the key assumption of a continuous latent math ability that the MAP Growth assessment is built on. Taken together, these strengths of the Saltus model make it particularly useful for this current study. Since the term "Saltus" was originally used due to the jumps in cognitive growth (Wilson, 1989), we can instead refer to the model as a structured confirmatory mixture IRT model. To date, no known applications of this model to CAT exist, making this research a novel application. The details of the parameterization of the structured confirmatory mixture IRT model used in this study can be found in Chapter 3.

2.5 Chapter Conclusion

A structured confirmatory mixture IRT model offers a unique opportunity to understand differences among early Kindergarten geometry subdomains. Other psychometric models reviewed in this section, while having useful strengths, also have limitations that make them less applicable in this context. This thesis aims to fill a gap in the psychometric literature by demonstrating the application of structured mixture IRT to computer adaptive data. The following chapter details the specific research questions and methods used for

this application.

CHAPTER 3

Methods

3.1 Research Questions

The goal of this application of structured mixture IRT is to explore differences in early kindergartener performance on the CCSS subdomains of geometry on the MAP Growth K-2 interim assessment. The specific research questions addressed in this analysis are:

1. Are there three substantively different groups of students in terms of geometry ability at kindergarten entry? In other words, is there evidence of three latent classes that represent low-performing, average-performing, and high-performing students?

2. Do these latent classes differ in their performance on the geometry subdomains? In other words: Do "analyze shapes" items differentiate students? Do "identify shapes" items?

3.2 Data

The sample for this study includes students who took the MAP Growth K-2 CCSS-aligned assessment in August or September of kindergarten in 2018. To study differences in geometry performance in different subdomains of interest, the sample is reduced to include only students who were administered items that are linked to the CCSS geometry standards in Kindergarten[1]. The final sample includes 125,668 students and available

[1]A small number of students, approximately 400, are removed because they are administered CCSS geometry linked items that are linked to standards at grades other than Kindergarten.

Table 3.1: Student demographics CCSS Geometry Samples

	CCSS Geometry Sample
N	125,668
Male	0.51
Race/Ethnicity	
Asian	0.04
Black	0.16
Hispanic	0.15
Other	0.18
White	0.47

demographics for the sample are presented in Table 3.1.

3.3 Structured Confirmatory Mixture IRT Model

In order to work with the current design of this mathematics assessment, a structured confirmatory mixture IRT model is used. As mentioned, one of the main benefits of this model is that it can be thought of as an extension of the Rasch model that is used in the original design of the assessment. Specifically, the structured confirmatory mixture IRT model that is used here incorporates latent classes and structured item group information (Jeon, 2018; Wilson, 1989). By extending the Rasch model, the model also estimates students' mathematics achievement on a continuous latent scale. However, by including latent classes, the it offers the opportunity to model latent heterogeneity which can serve to identify student's latent class (von Davier & Rost, 2006).

3.4 Formulation

The structured confirmatory mixture IRT model utilizes theory to set an a priori definition of the number of latent classes and item groups. For example, if there are two latent

classes representing average performing and below-average performing students at kindergarten entry ($C_j = h$, where $h = [1, 2]$). The item groups, $a = [1, 2]$, are defined as two subdomains of geometry. The probability of a correct response on item i for student j is conditional on overall latent geometry ability $\theta_{j(h)}$ and latent class assignment. It is modeled as,

$$P(Y_{ij} = 1|\theta_{j(h)}, C_j = h) = \frac{exp(\theta_{j(h)} - \beta_i + \tau_{hg} b_{ia})}{1 + exp(\theta_{j(h)} - \beta_i + \tau_{hg} b_{ia})}, \tag{3.1}$$

where $\theta_{j(h)}$ is overall latent geometry ability for student j in class h. Note that we are estimating the overall latent geometry ability, as opposed to overall math ability as in Equation 1.1, since only geometry items are included in this analysis. β_i represents the item location for item i. The item location can be interpreted similarly to the item difficulty in the Rasch model, but is referred to as the location, as oppposed to difficulty, since the overall item difficulty in this model is a combination of β_i and τ_{ha} (discussed more fully in Section 3.6). The τ_{ha} parameter is the unique aspect of this model. This parameter represents the effect of subdomain a on item responses for students in latent class h. Lastly, b_{ia} is an indicator variable denoting whether item i belongs to subdomain a. More details on the interpretation of this parameter are discussed below.

Latent ability is assumed to follow a normal distribution for each latent class ($\theta_{j(h)} \sim N(\mu_h, \sigma_h^2)$). Typically, the mean for one latent class is fixed to 0, making it the reference class. Variances can be fixed or estimated, depending on the goal of the analysis and identification. In this analysis, variances are freely estimated for each latent class.

3.5 Model Assumptions and Identification

Two sets of constraints are used in the model for model identification, as well as for aiding in the interpretation of parameters. The constraints used in this analysis follow those used in the original Saltus model (Mislevy & Wilson, 1996; Wilson, 1989). First, one latent class is set as the reference class. Typically, the mean of the ability estimate

for the reference class is set to zero. In this analysis, the mean of the reference class was set to the average score of kindergarten students at the fall test occasion after four weeks of instructional exposure (Thum & Kuhfeld, 2020), making the interpretation of the reference class, the "average" class. The analysis was conducted on the logit scale (without the linear transformation from Equation 1.2) and centered around the mean of the reference class.

Second, a set of constraints are set on the τ_{ha} parameter. One subdomain is considered the reference subdomain, such that the value of τ_{ha} for that subdomain is set to zero for all latent classes. Additionally, the value of τ_{ha} is set to zero for the reference class. Taken together, these constraints mean that for the reference class, it is assumed that no subdomains have higher or lower difficulty, and for the reference subdomain, that there are no differences between latent classes (Jeon, 2019). These assumptions are shown clearly when representing the τ_{ha} parameter in matrix form. Following the example in setting up Equation 3, with 2 latent classes ($h = [1, 2]$) and 2 item groups ($a = [1, 2]$), the τ_{ha} parameter matrix can be shown as,

$$\begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix}.$$

After incorporating the stated constraints, the τ_{ha} parameter matrix is,

$$\begin{pmatrix} 0 & 0 \\ 0 & \tau_{22} \end{pmatrix},$$

where τ_{22} can be interpreted as the difference in item difficulty for students in the non-reference class, compared to the reference class, on the subdomain of interest.

Now imagine that we have want to estimate three latent classes ($h = [1, 2, 3]$), below-average performing students, average performing students, and above-average performing students. If three latent classes are considered, the τ_{ha} parameter matrix can be shown as:

$$\begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \\ \tau_{31} & \tau_{32} \end{pmatrix}$$

and with the constraints as:

$$\begin{pmatrix} 0 & 0 \\ 0 & \tau_{22} \\ 0 & \tau_{32} \end{pmatrix}.$$

In this three-class case, τ_{ha} is estimated for the two non-reference latent classes and can be interpreted as the difference in item difficulty for that latent class, compared to the reference class.

3.6 Interpretation of τ_{ha} parameters

Jeon (2018) provides an in-depth discussion of how the τ_{ha} parameters can be interpreted as the difference in item difficulty for each latent class, compared to the reference group. Briefly, Equation 3.1 can be rewritten as:

$$logit(P(Y_{ij} = 1|\theta_j, C_j = h)) = \theta_{j(h)} - \underbrace{\beta_i + \tau_{ha} b_{ia}}_{\beta^*_{ih}}. \tag{3.2}$$

When Equation 3.2 is rewritten with β^*_{ih}, it gives:

$$logit(P(Y_{ij} = 1|\theta_j, C_j = h)) = \theta_{j(h)} - \beta^*_{ih}. \tag{3.3}$$

Note that Equation 3.3 is similar to that of Equation 1.1, with β^*_{ih} as the item difficulty for latent class h. Since latent class one is the reference class, $\beta^*_{i1} = \beta_i - \tau_{11} - \tau_{12} = \beta_i - 0 - 0 = \beta_i$, and the item location is equal to the item difficulty for this class. For latent class two, $\beta^*_{i2} = \beta_i - \tau_{21} - \tau_{22} = \beta_i - 0 - \tau_{22} = \beta_i - \tau_{22}$, and τ_{22} can be interpreted

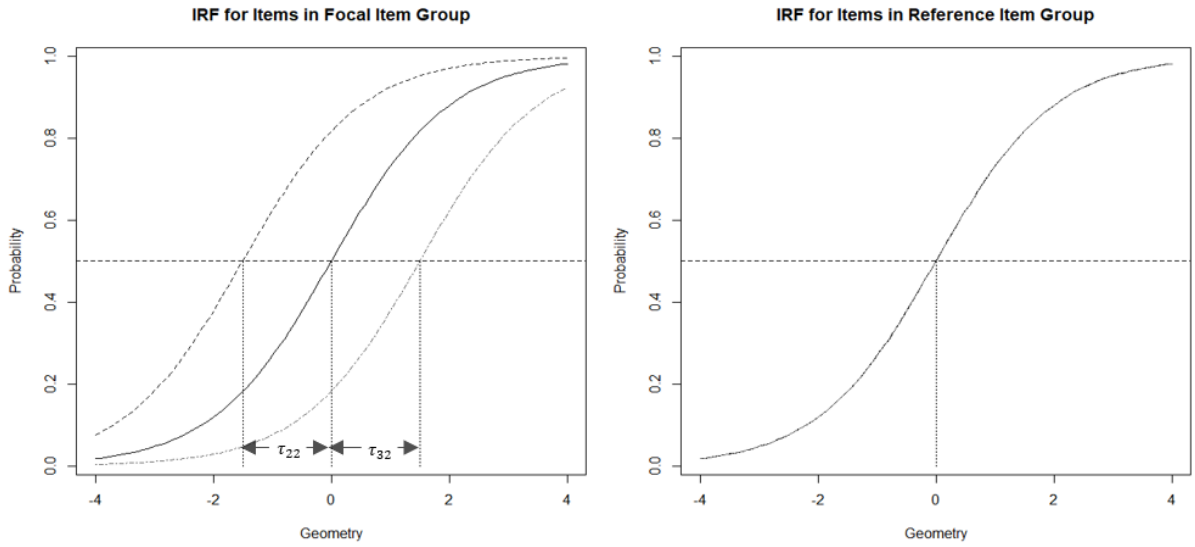

Figure 3.1: Item Response Function plots for an item with $\beta_i = 0$ in the focal item group (left) and in the reference item group (right), by latent class.

as the difference in item difficulty between latent class two and latent class one. The same interpretation also applies to latent class three, where $\beta^*_{i3} = \beta_i - \tau_{31} - \tau_{32} = \beta_i - 0 - \tau_{32} = \beta_i - \tau_{32}$. For more details, see Jeon (2018).

Figure 3.1 shows hypothetical item response functions (IRF; probability plot for a correct response at various levels of ability) for an item with $\beta_i = 0$ that is in the focal item group and one that is in the reference item group, by latent class in the three-class scenario. The left plot shows the IRF for items in the focal item group and the corresponding estimates of τ_{ha}. The solid black line is the IRF for the reference latent class. Since τ_{21} is fixed to zero, the item difficulty is equal to the item location. For latent class two, imagine that τ_{22} is estimated as 1.5, which shifts the IRF curve to the left for this class. In other words, the items in the focal item group are easier for students in latent class two. For latent class three, τ_{32} is estimated as -1.5, which shifts the IRF to the right. For this class, the items in the focal item group are more difficult. The plot on the right side of Figure 3.1 shows the IRF for items in the reference item group, where the item location (β^*_{ih}) is the same as the item difficulty (β_i) for all latent classes.

+

Table 3.2: Item group assumptions for Models 1 and 2

Model 1: Analyze Shapes	Model 2: Identify Shapes
1. There is a difference in item location on analyze, compare, create, and compose shapes items.	1. There is a difference in item location on identify and describe shapes items.
2. There is no difference on identify and describe shapes items.	2. There is no difference on analyze, compare, create, and compose shapes items.
3. No difference in item difficulty for any item group for the reference group.	3. No difference in item difficulty for any item group for the reference group.

3.7 Analytic Approach

Two sets of models were estimated, with a different focal item group for each model. In Model 1, the focal item group is analyze, compare, create, and compose shapes (and identify and describe shapes is the reference item group). In Model 2, identify and describe shapes is the focal item group. The assumptions that stem from the constraints on the τ_{ha} parameter matrix for each of these models are shown in Table 3.2. For each of these sets of models, a two-class and three-class models are estimated. To select the best model, two criteria are used. First, relative fit statistics are used to compare the fit of the two-class and three-class models to a one-class model (that is the Rasch model). Second, the tenability of the assumptions outlined in Table 3.2 is evaluated.

All models are estimated in Mplus version 8 (Muthén & Muthén, 2017) using maximum likelihood estimation. Latent class posterior probabilities are estimated, and individual class membership is based on the most likely class. The item location parameters are fixed at the values based on the original MAP Growth K-2 item calibrations. The mean of the reference group is set using the average value for Kindergarten entry, such that

the reference class can be considered "average" students in both the two- and three-class models. The mean for the non-reference class and the variances for both classes are estimated.

CHAPTER 4

Results

4.1 Item Descriptive Statistics

Before fitting the structured confirmatory mixture IRT model, it is important to understand the distribution of items across the CCSS geometry domain. Table 4.1 presents item statistics for all CCSS geometry items, as well as broken down by subdomain. Overall, there are 135 items linked to CCSS kindergarten geometry domain in the item pool. 92 of those items are identify shapes items and 43 are the analyze shapes items. Due to the adaptive nature of the test and blueprint specifications, students differ in the number of items they see by subdomain (but are more similar on overall geometry, as this is specified in the blueprint). On average, students see more Identify Shapes items (5.77 items) compared to Analyze Shapes items (2.34 items). One limitation is that a small number of students were administered items only from one subdomain, and when this is the case, students are typically administered identify items. This is due to the blueprint specifications only existing at the domain level, and not the subdomain level. The implication of this design is that individual student classifications should not be utilized at this time. However, we are still able to explore the existence of latent classes and differentiation by domain, at the aggregate level.

The average proportions correct are equal in each subdomain and across the full domain. This is expected due to the nature of adaptive testing, where the items are selected based on the item difficulty associated with an expected probability of 50% for a student to respond correctly to the item. However, there are clear differences in the spread of the proportion correct by item subdomain. Specifically, the standard deviation

Table 4.1: Number of items and proportion correct for all CCSS geometry and by subdomain

Subdomain	# of items	Average # per student	Average Proportion Correct	SD Proportion Correct	Average Item RIT	SD Item RIT
Identify and describe shapes	92	5.77	0.55	0.24	139.07	7.70
Analyze, compare, create, and compose shapes	43	2.34	0.55	0.38	138.38	12.03
All Geometry	135	8.11	0.55	0.19	138.76	10.16

on the proportion correct for analyze shapes items is quite large (0.38), compared to identify shapes (0.24). Similarly, the average item RIT (item difficulty) is similar across the subdomains but the standard deviation is largest for the analyze items. These differences suggest that there are more differences among students on this subdomain of geometry.

4.2 Number of Latent Classes

Both models were estimated with two- and three-class solutions. Table 4.2 presents AIC, BIC, and adjusted BIC statistics for the two- and three-class models, as well as the one-class model. For both models, the three-class solution shows relatively the best fit. The two-class solutions show better fit compared to the one class model for both models. This suggests the existence of latent classes in terms of students' geometry performance on geometry subdomains at Kindergarten entry. Since the three-class solution is the best fitting solution for both models, results from the three-class models are selected.

Table 4.2: Relative fit statistics for 1, 2, and 3 class analysis for Models 1 and 2.

		Model 1		Model 2	
	1 class	2 class	3 class	2 class	3 class
AIC	1491131.46	1490436.90	1480893.20	1490438.08	1480585.32
BIC	1491141.20	1490466.13	1480980.87	1490467.30	1480672.94
Adjusted BIC	1491138.02	1490456.59	1480952.27	1490457.77	1480644.39

4.3 Overall Probabilities

Before interpreting individual parameter estimates for the structured confirmatory mixture IRT model, it is useful to look at the overall probability of a correct response for the average student in each latent class. This provides context for understanding differences between the two models and for viewing the tenability of the model assumptions.

Figure 4.1 shows the probability correct plotted against items of increasing item difficulty for both geometry subdomains. Note that the probability plots are essentially identical for both models. Class 1 (shown as circles) is the average class that is fixed as the reference group. For this class, the overall probability of a correct response is the same across both item groups (by design). Class 2 (shown as triangles) is an interesting class because the average student in this class has a higher probability of a correct response on analyze shapes items compared to class 1, but a lower probability on identify shapes items. Class 3 (shown as diamonds) has the lowest probability of a correct response on analyze shapes items, but no difference compared to the reference class on identify shapes items.

4.4 Tenability of Model Assumptions

To determine which model is more appropriate, we can assess the tenability of the model assumptions presented in Table 3.2 by looking at the overall probabilities. In Model 1, the focal item group is analyze shapes items and the reference item group is the identify

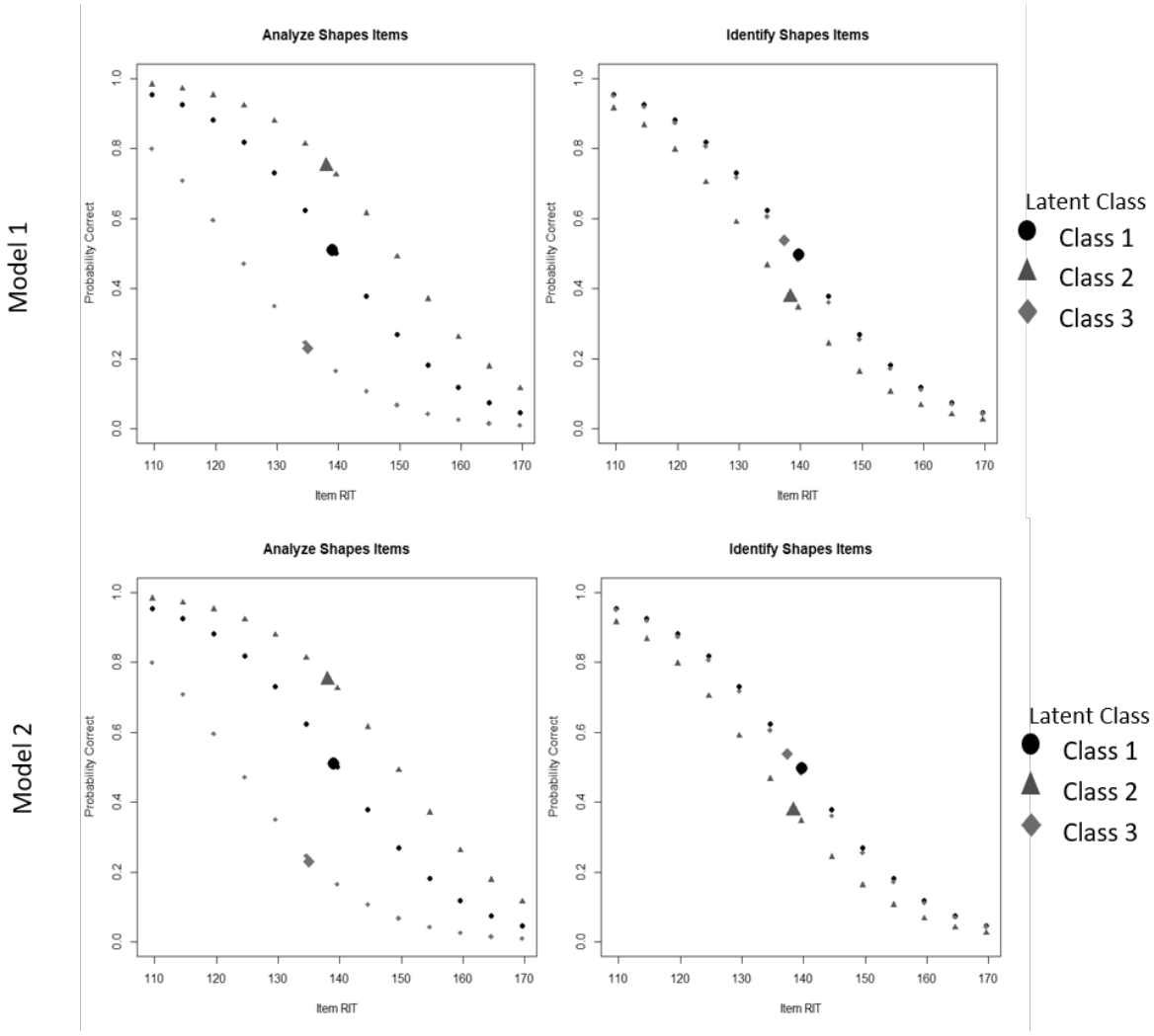

Figure 4.1: Probability of a correct response plotted against item location on the RIT scale for Model 1 (top row) and Model 2 (bottom row). Class 1 is shown as circles, class 2 as triangles, and class 3 as diamonds. The large shape shows the average item RIT for students in each latent class, on each subdomain.

shapes items. This means that Model 1 assumes that there is no difference on identify shapes items. In Model 2, the focal item group is the identify shapes items and the reference group is the analyze shapes items. Model 2 assumes that there is no difference on analyze shapes items. Since there is much less differentiation between latent classes on the identify shapes items, the assumptions from Model 1 seem more tenable than those of Model 2. For this reason, in addition to the fact that the overall story is the same across models, results from Model 1 will be emphasized while differences in Model 2 will be briefly discussed.

4.5 Latent Class Proportions

Counts and proportions for class assignment are based on the estimated model parameters. In Model 1, 51% of students are classified in class 1 (the average class), 25% in class two, and 24% in class 3. Almost all students who are classified in class 2 or 3 in Model 1 are classified in the same class in Model 2. However, students who are classified as class 1 are classified more often in classes 2 and 3 in Model 2. One possible explanation is that the impact of the blueprint design makes it difficult to classify students when the focus is on analyze items. For this reason, individual student classifications should not be used at this time.

4.6 Latent Class Parameter Estimates

Table 4.3 presents parameter estimates for the latent class distributions on both the latent logit scale and the RIT scale. Parameter estimates are different across models, and at first glance may seem to tell a different story. However, recall that when the parameters are interpreted together, the probabilities of correct responses by latent class are the same across models.

Table 4.3: Estimated mean and variance for geometry ability on the latent logit scale, τ_{ha}, and mean and standard deviation on the RIT scale, by latent class.

		Model 1			Model 1		
		Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Latent logit scale							
	$\hat{\mu}_h$	0.00	-0.64	-0.07	0.00	0.85	-1.36
		–	(0.02)	(0.02)	–	(0.04)	(0.04)
	$\hat{\sigma}_h^2$	2.10	0.92	0.88	2.64	1.07	0.98
		(0.11)	(0.03)	(0.03)	(0.07)	(0.03)	(0.03)
	$\hat{\tau}_{ha}$	0.00	1.60	-1.54	0.00	-1.31	1.32
		–	(0.05)	(0.04)	–	(0.04)	(0.04)
RIT scale							
	mean	139.56	133.19	138.85	139.56	148.03	125.97
	sd	14.48	9.60	9.38	16.26	10.34	9.91

4.6.1 Model 1: Analyze Shapes Model

Both classes 2 and 3 have statistically significant differences in estimated average ability compared to class 1. The difference for class 2 is moderate in size ($\hat{\mu}_2 = -0.64, SE = 0.02$) and is equal to less than 6 points on the RIT scale. This small difference means that the differences between class 2 and class 1 are not fully due to the analyze shapes domain. Specifically, because there is only one other subdomain, we know that the difference in average ability is coming from the identify shapes domain for class 2. For class 3, the difference is statistically significant, but very small ($\hat{\mu}_3 = -0.07, SE = 0.02$), equal to approximately a point difference on the RIT scale. The lack of difference in estimated ability suggests that the differences between class 3 and class 1 are captured in the analyze shapes items. The left histogram in Figure 4.2 shows the distribution of geometry ability for Model 1, with dotted lines for the mean of each class. The larger difference for class 2 is easily visible in this plot.

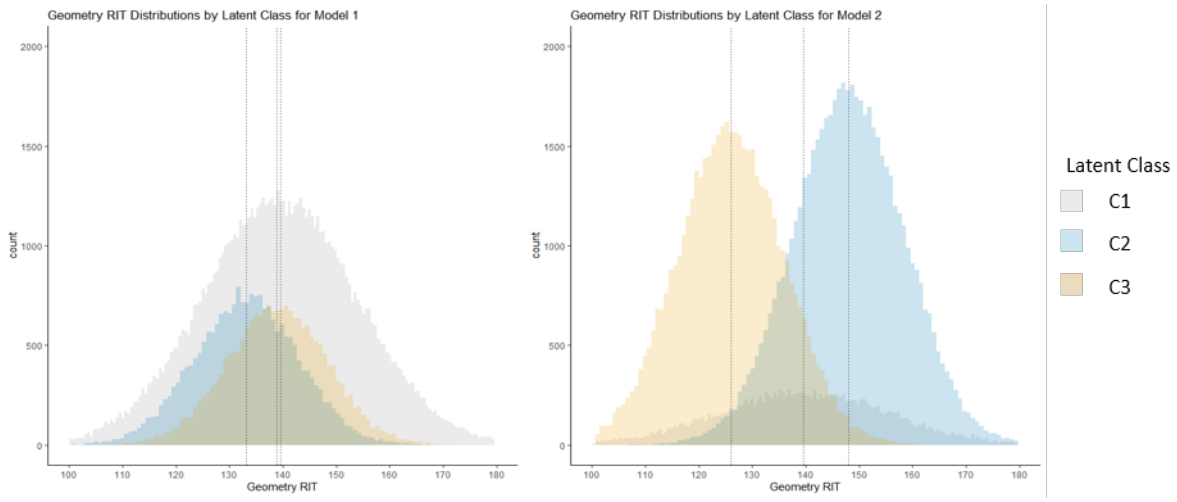

Figure 4.2: Histograms of geometry ability distributions for Model 1 (left) and Model 2 (right). Class 1 is shown in gray, class 2 in blue, and class 3 in yellow.

The estimated variance components for each class are also presented in Table 4.3. The class with the largest spread is class 1 ($\hat{\sigma}_1^2 = 2.10, SE = 0.11$), suggesting that there are large differences in overall ability within this class, but that these differences are attributed to all geometry, not to a specific subdomain. The estimated variance components for classes 2 and 3 are similar ($\hat{\sigma}_2^2 = 0.92, SE = 0.03; \hat{\sigma}_3^2 = 0.88, SE = 0.03$), suggesting that there are moderate differences in overall ability within both classes, even after for accounting for the performance on analyze shapes items.

The final parameter to interpret for this model is the estimate of differences in item difficulty on analyze shapes items. For class 2, analyze shapes items are easier compared to the reference class ($\hat{\tau}_2 = 1.60, SE = 0.05$). The left plot of Figure 4.3 shows the IRF plot for analyze shapes items in Model 1 for a sample item with the item location equal to zero ($\beta_i = 0$). The solid black line represents class 1 as the reference class. The dotted line to the left represents the probability of a correct response for class 2 at different levels of geometry ability (θ_j), and shows how students in this class can have a lower level of estimated ability with an associated probability of a correct response equal to 0.50. For class 3, the opposite is true. For students in this class, analyze shapes items are more difficult compared to the reference class ($\hat{\tau}_3 = -1.54, SE = 0.04$). The dashed, gray line to

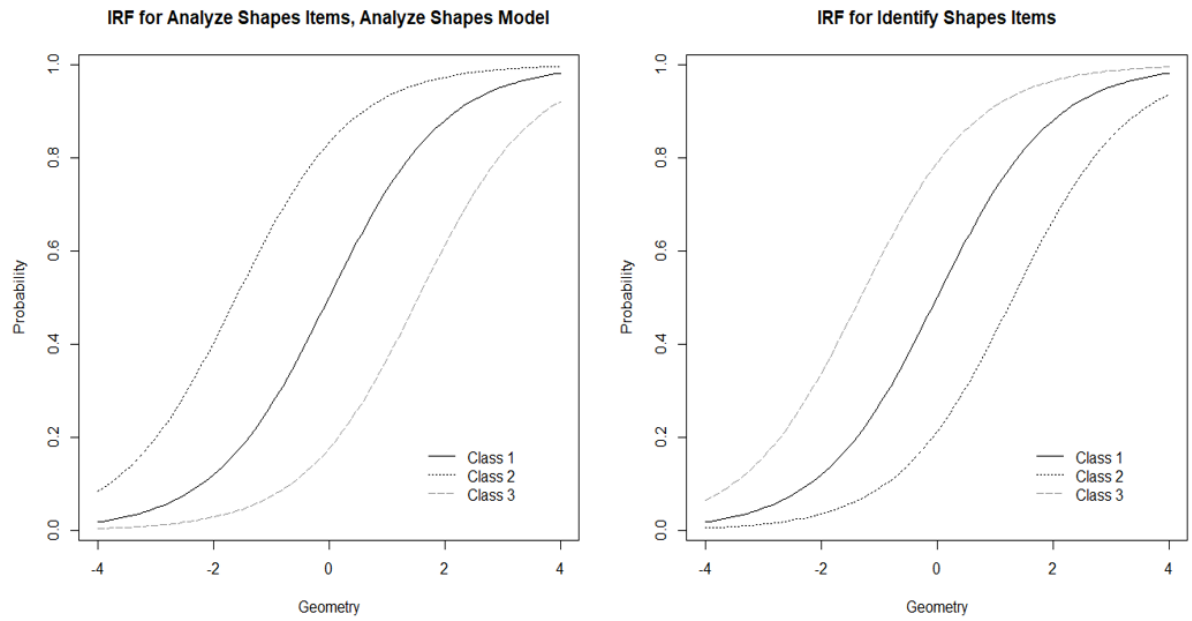

Figure 4.3: Item Response Function (IRF) plots for focal item groups Model 1 (left) and Model 2 (right). Solid line represents reference class, dashed line is class 2, and dotted line is class 3.

the right of the reference class in the left plot of Figure 8 shows how these students need to have a higher level of estimated ability to have a 50% chance of responding correctly to these items.

Taken together, we get the probabilities shown in Figure 4.1. The large and positive difference in item difficulty for class 2 means that, despite the slightly lower mean geometry ability, students in this class typically do better on analyze shapes items. In fact, the slightly lower mean tells us that there is also a difference for these students on identify shapes items, as well see in the slightly lower probability of a correct response on those items for class 2. For class 3, the large and negative difference in item difficulty means that students in this class typically perform worse on the analyze items. Since there is no real difference in mean geometry ability between these classes, we know that there is no real difference on identify shapes items for class 3.

4.6.2 Model 2: Identify Shapes Model

Results for Model 2 are foreshadowed by the results in Model 1, and for that reason, they are only briefly discussed here. Since we know from Model 1 that there are differences in analyze items across latent classes, we expect that these differences will show up in the mean geometry ability when not accounted for in the item difficulty. This expectation proves true for both latent classes. For class 2, the estimated mean is slightly higher ($\hat{\mu}_2 = 0.85, SE = 0.04$), capturing the fact that the analyze items are somewhat easier for these students. In class 3, the estimated mean is much lower ($\hat{\mu}_3 = -1.36, SE = 0.04$), capturing the fact that analyze items are much more difficult for this class. We also expect to see estimates for differences in item difficulty for identify shapes items that reflect the fact that there are differences in this domain for class 2 but not for class 3. For class 2, the identify shapes items are more difficult ($\hat{\tau}_2 = -1.31, SE = 0.04$). The large estimate difference serves to balance out the higher estimated mean, bringing the overall probability for this class on these items to a somewhat lower level compared to class 1, (as shown in Figure 4.1). For class 3, the fact that the identify items are estimated to be easier ($\hat{\tau}_2 = 1.32, SE = 0.04$), in fact balances out the large mean difference that stems from the analyze items. In other words, there are no differences on average for class 3 in overall probability on identify shapes items, compared to class 1.

4.7 Item Statistics by Latent Class

With an understanding of differences across these three latent classes, we can return to the item statistics presented in Table 4.1, but now broken down by latent class. Table 4.4 presents the average proportion correct for each latent class on the different subdomains and overall. There are small differences in proportion correct between students in classes 1 and 3 on identify shapes items, while students in class 2 shows a smaller proportion correct on these items. On the other hand, class 2 does much better on analyze shapes items, while class 3 students do quite poorly on this subdomain. None of

Table 4.4: Average proportion correct by latent class, for subdomains and total items for Models 1 and 2.

		Model 1		Model 2	
CCSS Group	Class	Avg Prop	SD	Avg Prop	SD
Identify and Describe Shapes	C1	0.57	0.25	0.64	0.33
	C2	0.41	0.19	0.48	0.21
	C3	0.62	0.19	0.60	0.2
Analyze, Compare, Create, and Compose Shapes	C1	0.55	0.34	0.60	0.31
	C2	0.90	0.20	0.79	0.28
	C3	0.15	0.22	0.26	0.29
Total	C1	0.57	0.20	0.64	0.26
	C2	0.55	0.15	0.56	0.17
	C3	0.49	0.15	0.51	0.17

these large differences are clear when looking at proportions correct on the total items. Without exploring latent classes defined by subdomains, we might miss the more nuanced differences in geometry performance between these three latent classes.

We can also look at item difficulty and number of items by latent class to understand whether there are differences in how items are administered to students across latent classes. Table 4.5 shows the average number of items administered to students by specific CCSS standard within each subdomain. For all standards, the differences in average item RIT are small across latent classes. The largest differences are in the analyze shapes domain, where standard B4 has a difference of about 4 RIT points, on average. Standard B6 has a larger difference but students in all latent classes are seeing very few of these items. The number of items per standard is similar across each class on each subdomain. In the identify shapes domain, students in all classes are seeing the most items for standard A2 and a moderate number for standard A1. This means that students in class 2 mostly differ on these two standards. In the analyze shapes domain, students are primarily seeing items from standard B4. This means that differentiation among latent

Table 4.5: Average number of items and average item RIT, by CCSS and latent class.

CCSS Group	CCSS	Class	Model 1			Model 2		
			Avg # of items	Avg Item RIT	Item RIT SD	Avg # of items	Avg Item RIT	Item RIT SD
Identify and Describe Shapes	A1	C1	1.53	139.54	8.42	1.27	142.41	10.77
		C2	1.53	138.18	7.93	1.58	138.69	7.83
		C3	1.57	137.63	7.15	1.60	138.25	7.29
	A2	C1	3.76	138.14	7.75	3.03	139.05	9.15
		C2	3.87	137.14	7.58	3.90	137.67	7.49
		C3	4.01	136.87	7.40	4.04	137.45	7.38
	A3	C1	0.25	163.75	4.62	0.56	164.34	4.65
		C2	0.19	162.70	4.09	0.19	162.96	4.34
		C3	0.06	161.41	3.69	0.10	162.66	4.44
Analyze, Compare, Create, and Compose Shapes	B4	C1	2.75	139.19	12.08	3.41	139.75	13.56
		C2	2.87	138.38	11.77	2.66	139.33	15.44
		C3	3.12	135.53	11.28	2.81	136.75	11.31
	B6	C1	0.08	139.11	14.76	0.10	148.38	15.44
		C2	0.11	136.48	13.49	0.09	137.00	13.75
		C3	0.07	131.87	9.06	0.07	132.45	9.84

classes comes from items measuring this particular standard.

CHAPTER 5

Discussion

This thesis has presented an application of structured confirmatory mixture IRT model to adaptive testing and demonstrated its potential in the context of supporting geometry instruction at kindergarten entry. Evidence from relative fit statistics supports the existence of three latent classes of student geometry performance. However, these classes do not necessarily represent low, medium, and high performing students, but instead present a more nuanced picture of differences among early Kindergarteners. Both subdomains of geometry differentiate among latent classes, but the identify shapes domain only differentiates class 2, whereas the analyze shapes domain differentiates all three.

Since the assessment blueprint does not specify the number of items per subdomain to administer to each student, individual classifications for students are not usable. However, broad conclusions regarding differences among students on the subdomains can be discussed. Specifically, one class of students (class 3) is comparable with average students on identify shapes items, but struggle with analyze shapes items. This area of relative strength for students in this class can be mostly attributed to standards related to naming shapes and describing relative positions (see Table 4.5 for item numbers and Table 1.1 for standards details). The area of relative weakness for students in this class relates to comparing and analyzing two- and three-dimensional shapes. Class 2, on the other hand, shows an area of strength compared to average students on these items comparing and analyzing two- and three-dimensional shapes, yet struggle somewhat with naming shapes and describing relative positions. Taken together, these differences across latent classes suggest that the acquisition of early geometry skills does not happen in a linear fashion for all students. However, to more fully understand the implications of these

conclusions, some additional areas of research are needed.

The first area of future research relates to the construct definition and subdomains. First, a clear understanding of how items map to the construct and subdomains is essential. In the context of MAP Growth K-2, the items are linked retroactively to only a single CCSS standard, rather than being created to measure one specific area. It is possible, even likely, that some items could be considered to measure more than one specific standard. If this is the case, understanding and even modeling these cross-domain items would provide further insight into the nature of early geometry skill acquisition. Relatedly, future work should consider how the definition of the construct represented by θ_{jh} shifts when accounting for differences in difficulty for some subdomains. This would be an essential discussion when considering comparing results from this model to other test results.

The second area of future research relates to individual classifications. In contexts where individual classifications are appropriate, exploring future performance based on these early kindergarten latent classes can provide insight into whether, and for how long, these differences remain after exposure to formal instruction. Another area of research on individual classifications could incorporate covariates for predicting latent class membership as demonstrated by Jeon (2018), allowing for a better understanding of the students within each class.

REFERENCES

Abry, T., Latham, S., Bassok, D., & LoCasale-Crouch, J. (2015). Preschool and kindergarten teachers' beliefs about early school competencies: Misalignment matters for kindergarten adjustment. *Early Childhood Research Quarterly*, *31*, 78–88.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*(4), 381–409. Retrieved 2017-07-11, from `http://journals.sagepub.com/doi/abs/10.3102/10769986026004381`

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

De Boeck, P., & Wilson, M. (2014). 12 multidimensional explanatory item response modeling. *Handbook of item response theory modeling: Applications to typical performance assessment*, 252.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343–368.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359–374.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement*, *44*(4), 325–340.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. L. . M. J. Gierl (Ed.), *Cognitive diagnostic assessment for education: Theory and applications* (p. 242-274). New York, NY: Cambridge University Press.

Gnaldi, M., Bacci, S., & Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, *10*(1), 53–70.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* (Unpublished doctoral dissertation). ProQuest Information & Learning.

Jeon, M. (2018). A constrained confirmatory mixture irt model: Extensions and estimation of the saltus model using mplus. *The Quantitative Methods for Psychology*, *14*, 120–136.

Jeon, M. (2019). A specialized confirmatory mixture irt modeling approach for multidimensional tests. *Psychological Test and Assessment Modeling*, *61*(1), 91–123.

Jeon, M., Draney, K., & Wilson, M. (2015). A general Saltus LLTM-R for cognitive assessments. In *Quantitative psychology research* (pp. 73–90). Springer.

Kuhfeld, M., Soland, J., Pitts, C., & Burchinal, M. (2020). Trends in children's academic skills at school entry: 2010 to 2017. *Educational Researcher*, *49*(6), 403–414.

Linden, W. J., van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Springer.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from irt-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika*, *61*(1), 41–71.

Muthén, B., & Muthén, L. (2017). *Mplus*. Chapman and Hall/CRC.

Newcombe, N. S., & Frick, A. (2010). Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education*, *4*(3), 102–111.

NWEA. (2019). *Map growth technical report* (Tech. Rep.). Portland, OR.

Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*(350), 351–356.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Reardon, S. F., & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *Aera Open*, *2*(3), 2332858416657343.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219–262.

Smit, J., Kelderman, H., Flier, H., et al. (2000). The mixed birnbaum model: Estimation using collateral information.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement*, *44*(4), 313–324.

Thum, Y. M., & Kuhfeld, M. (2020). *Nwea 2020 map growth achievement status and growth norms for students and schools* (Tech. Rep.). Portland, OR.

Van der Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing* (pp. 3–30). Springer.

Verdine, B. N., Irwin, C. M., Golinkoff, R. M., & Hirsh-Pasek, K. (2014). Contributions of executive function and spatial skills to preschool mathematics achievement. *Journal of experimental child psychology*, *126*, 37–51.

von Davier, M., & Rost, J. (2006). *Mixture distribution item response models* (Vol. 26). Elsevier.

von Davier, M., Xu, X., & Carstensen, C. H. (2009). Using the general diagnostic model to measure learning and change in a longitudinal large-scale assessment. *ETS Research Report Series*, *2009*(2), i–22.

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "are schools the great equalizer?". *Sociology of Education*, *91*(4), 323–357.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past

is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352–360.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276.

Wolf, S., Magnuson, K. A., & Kimbro, R. T. (2017). Family poverty and neighborhood poverty: Links with children's school readiness before and after the great recession. *Children and Youth Services Review*, *79*, 368–384.