2021

Moving Past ‘One Size Fits All’: Developing a Trajectory Deviance Index for Dynamic Measurement Modeling

Yixiao Dong
University of Denver

Moving Past ‘One Size Fits All’: Developing a Trajectory Deviance Index for Dynamic Measurement Modeling

---

A Dissertation

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Yixiao Dong

June 2021

Adviser: Denis Dumas

Author: Yixiao Dong
Title: Moving Past ‘One Size Fits All’: Developing a Trajectory Deviance Index for Dynamic Measurement Modeling
Adviser: Denis Dumas
Degree Date: June 2021

Abstract

Dynamic Measurement Modeling (DMM) is a recently developed measurement framework for gauging developing constructs (e.g., learning capacity) that conventional single-timepoint tests cannot assess. Like most measurement models, overall model fit indices of DMM do not indicate the measurement appropriateness for each included student. For this reason, other measurement modeling paradigms (e.g., Item-Response Theory; IRT) utilize person-fit or model appropriateness statistics to indicate whether a measurement model appropriately describes the data from each individual student. However, within the extant DMM framework, no statistical index has yet been developed for this purpose. Thus, the current project advanced a person-specific DMM Trajectory Deviance Index (TDI) that captures the aberrance of an individual’s growth from the model-implied trajectory. Two simulation studies were conducted to examine and compare the distributional properties and effectiveness of four TDI candidates with different formulations. Consequently, the best functioning one was determined as the final formulation of the TDI. The data generation model was based on the parameter estimates from the Technology-enhanced, Research-based, Instruction, Assessment, and professional Development (TRIAD) cluster-randomized experiment data, which contains seven waves of mathematics test scores for students from pre-school through Grade 5. Besides the simulation work, an empirical study was also conducted to demonstrate the uses of the developed TDI within those real-world data. The results indicated that bilingual status was significantly related to the deviance of growth in early mathematics, whereas the other examined factors (i.e., intervention, age, gender, special education status, Socioeconomic status) were not. Incorporating TDI into DMM analysis strengthened the validity of score use and interpretation, and offered a quantitative means of determining which students in the dataset were not adequately served by the dynamic measurement model.

Acknowledgments

The present study uses the data from the TRIAD research project. That original research was supported by the Institute of Education Sciences, U.S. Department of Education, through grants R305K05157 and R305A110188 and also by the National Science Foundation, through grants ESI-9730804 and REC-0228440. The author declares that he has no conflicts of interest. This research has followed all ethical guidelines of the American Psychological Association and was approved by the Institutional Review Board at the University of Denver. Correspondence concerning this study should be addressed to Yixiao Dong, Department of Research Methods and Information Science, University of Denver, Denver, CO, 80208. Email: yixiao.dong@du.edu.

I first wish to express my appreciation to my dissertation committee (Denis Dumas, Kathy Green, Douglas Clements, Julie Sarama, and Daniel McNeish), without whom this dissertation would not have been possible. Most of all, I would like to thank my mentor, Denis, for his efforts to cultivate me through the years and help me launch my academic career. I also want to thank my wonderful lab-PIs, Doug and Julie, for their support and contributions to my development as a researcher.

And my heartfelt thanks to my family, especially my wife Shuyuan, who even allowed me to work on my writing during our wedding night and the birth week of our child. Finally, but not least, I want to thank my toddler Rosely for being collaborative in the process of producing this dissertation.

Table of Contents

List of Figures

Chapter One: Introduction

Work in the field of dynamic testing has suggested interesting paradigms and ideas as well as promising findings. The question is whether this potential can be realized in a branch of psychological testing characterized by consistently converging results and techniques that provide information over and above the data collected by conventional tests. We believe that dynamic testing will ultimately meet these challenges and will prove to be a valuable resource to the psychological profession and to the world.

---Robert J. Sternberg & Elena L. Grigorenko, *Dynamic Testing* (2002)

Educational researchers care about students' potential to grow and learn as well as how much they currently know (Dweck, 2015; Vygotsky, 1931/1997). In the current educational system, single-timepoint tests (also termed static tests, Dumas, McNeish & Greene, 2020; Sternberg & Grigorenko, 2002) are often capable of measuring students' current ability but not their future learning capacity. Dynamic assessment (DA), or a measurement diagram featuring multiple testing occasions with integrated instructions (Feuerstein, 1979; Tzuriel, 2001), is an essential methodology to capture students' learning capacity or learning potential. Since Feuerstein first formalized the DA theoretical framework after World War II, DA associated methods have been applied in empirical studies where evidence was found to support that theorizing (e.g., Feuerstein et al., 1987; Swanson, 1995; Sternberg et al., 2002). However, as indicated by Sternberg and Grigorenko (2002) in the opening quotation, although DA research demonstrated promising findings to researchers, it had not fully developed as a widely used methodological framework. Major challenges occur when applying DA theories to educational research and practices. For example, following Feuerstein's conceptualization, clinician training and instructional interventions are required, which could make the operational costs of DA too high for conducting large-scale studies. Additionally, there was a lack of a robust statistical framework to achieve the important goals of DA (e.g., quantifying capacity in a reliable way; Dumas et al., 2020). These issues were not well-addressed until Dumas and McNeish (2017) further developed the DA theories and concepts. They termed their formal statistical modeling framework, *Dynamic Measurement Modeling* (DMM), to measure learning capacity by producing learner-specific asymptotes to a growth trajectory.

Significance of DMM to Educational Research, Evaluation and Measurement Practices

DMM integrated the DA theories and goals into modern psychometric and statistical approaches (Dumas et al., 2020), which lay the methodological foundation for educational psychologists to estimate students' learning potential and conduct investigations of related topics. Since Dumas and McNeish published their initial DMM work in 2017 (Dumas & McNeish, 2017; McNeish & Dumas, 2017), empirical evidence associated with the effectiveness of DMM in research and measurement practices has continued to accumulate. Up to now, DMM has been applied to the research of various educational and cognitive constructs such as mathematics abilities (Dumas & McNeish, 2017; Dumas et al., 2020; Dumas, McNeish, Sarama & Clements, 2019; McNeish & Dumas, 2020), medicine related knowledge (Dumas, McNeish, Schreiber-Gregory & Durning, 2019), and reading and verbal ability (Dumas & McNeish, 2018; McNeish et al., 2019). More importantly, this theoretical-psychometric paradigm reveals new paths to deal with chronic educational problems around the world. Two examples are illustrated here.

Sustained Achievement Gaps

The achievement gap has been one of the most discussed and studied topics in U.S. education (Ladson-Billings, 2006) and is also extensively researched around the world (e.g., U.K., von Stumm, 2017; France, Crouzevialle & Darnon, 2019; Japan and Korea, Holloway et al., 2016). The term generally refers to the disparities in various measures of educational performance among subgroups of students, and these measures are usually single-timepoint standardized tests. However, the static measurement practices can only assess how much students have learned in a specific domain at the time of assessment rather than how much students may be capable of learning in the future (Dumas & McNeish, 2017; Sternberg et al., 2002). Given that students from many disadvantaged achievement groups (e.g., lower socioeconomic-status groups, African American and Hispanic groups in the U.S.) are historically marginalized, they are less likely to receive necessary instruction and other learning opportunities to help them catch up. The single-timepoint achievement scores are commonly used to *predict* their future performance or as a reference to distribute educational resources. Consequently, certain types of achievement gaps (e.g., racial gaps in the U.S.) are persistent, and the advance of social justice in the U.S is also hindered. As early as a century ago, DuBois (1920) already criticized conventional testing practices saying they would *create* students' futures, but the issue has never been substantially resolved.

In contrast, DMM, featuring dynamic assessment practices, can provide researchers with much richer growth information for each student, including estimates of learning capacity, learning rate, and forms of growth trajectory. The capacity (learning potential) estimates, in particular, may serve as important evidence for evaluating students' learning along with their current achievement levels. Notably, the capacity scores from DMM have been empirically demonstrated to be much less impacted by students' demographics (e.g., socioeconomic status, race/ethnicity, and gender, Dumas & McNeish, 2017) and high-quality instructional experience (Building Blocks intervention, Dumas, McNeish, Sarama & Clements, 2019) than conventional achievement scores. Therefore, students with disadvantaged backgrounds and learning experiences may still show high learning capacity in the DMM context. In this way, DMM capacity scores could be a unique source for making equitable educational decisions as well as improving the consequential validity (fairness) of educational measurement.

High-stakes Testing Struggles: Academic Involution and Examination Hell

Compared to the world-wide achievement gap issue, academic involution is more specific to Asian countries at the current period of time. American anthropologist Alexander Goldenweiser first coined the term "involution" as a culture or ecology that does not expand its economy but only develops its internal complexity and inefficiency (Hui, 2009). Geertz (1963) later borrowed this term to describe agricultural involution in Indonesia, and recently educational researchers have started to be concerned about the intensifying involution in academic settings (Kapur & Perry, 2014). Students from many Asian countries (e.g., China, India, Japan, and South Korea) spend a large amount of time doing repetitive practices in their daily learning. Such inefficient learning may not expand students' knowledge but only may result in potential increases in high-stakes test scores, which is a manifestation of educational involution. Meanwhile, these students usually need to suffer a variety of mental and psychical challenges to perform better in the high-stakes tests (c.f., *examination hell* in Japan and South Korea, Haberman, 1988; Lee, 2003). It is also common to see item content beyond the curriculum and teaching syllabus included in those tests to differentiate students with various levels of abilities further: a strategy used to create variance among students that may unfortunately reduce the validity of the measures.

From the perspective of measurement and educational evaluation, one important reason for the problem is the single-timepoint test score (e.g., college examination tests) as the sole criteria for distributing educational resources or making other critical life decisions. The large student population in East Asian countries does not allow for the individual application evaluation for college admissions used by the western education systems. In this case, DMM estimates could be used for a large population and serve as a second criteria for evaluating students' ability. Within the DMM conceptualization, all the repetitive practices simply reflected the process or effort of getting closer to the asymptotic capacity for each student (see *Figure 1* below). Because DMM is a reliable tool for measuring learning capacity as a latent variable (Dumas et al., 2020; Dumas & McNeish, 2017; McNeish & Dumas, 2018), students from those countries may therefore not need to spend as much time on inefficient learning in order to make their capacity quantifyable.

Developing a Trajectory Deviance Index for Dynamic Measurement Modeling

The establishment of DMM appears to be a positive answer to the anticipation of dynamic testing in the opening quotation (Sternberg & Grigorenko, 2002), and further steps may concentrate on how to make DMM a better and more valuable resource to the psychological profession and to the world. In addition to a wide range of application studies using DMM, methodological advances are simultaneously achieved (e.g., the development of a DMM conditional reliability index, McNeish & Dumas, 2018; incorporating seasonal learning loss within DMM, McNeish & Dumas, 2020). Refining and improving this newly invented methodological framework is a continuing task.

Problem Statement

The unique capacity scores from DMM have been estimated through nonlinear mixed-effects techniques. Several types of growth trajectory models (e.g., Michaelis-Menten, exponential, logistic) have also been applied to describe the data, and the best-fitting functional forms have typically been identified by comparing model fit indices (McNeish & Dumas, 2017). Like most measurement models, model fit indices for DMM only indicate the overall measurement appropriateness, which means a chosen function represents the overall growth curve (e.g., Michaelis-Menten trajectory for TRIAD data) rather than the growth of every student. Consequently, parameter estimates from DMM could be inaccurate for misfitting cases, which is a validity threat to DMM.

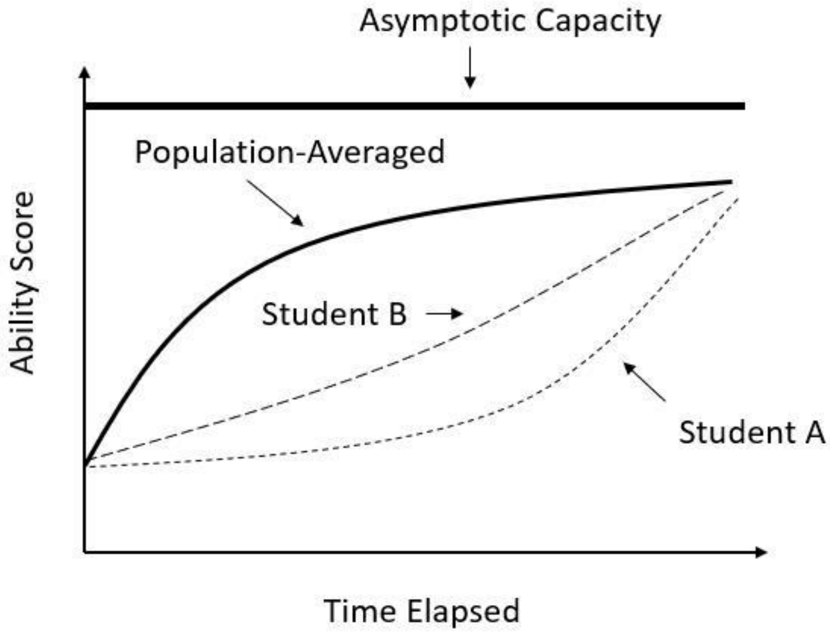


Figure 1. An Overall Growth Trajectory and Specific Trajectories for Potential Misfitting Students

Figure 1 provides an example of an overall growth pattern and student-specific trajectories for two exemplar aberrant cases. The overall curve featured a pattern that students' ability grew faster in early timepoints and leveled off with time elapsed. Such an inverted-J shape curve and its related function have been often applied to describe growth data in DMM (Dumas & McNeish, 2017; Dumas, McNeish, Sarama & Clements, 2019; McNeish et al., 2019; McNeish & Dumas, 2017), and the asymptote against the curve was conceptualized as the learning capacity of the ability growth. However, students A and B appeared to grow in very different trajectories than the typical trajectory that most students followed. It should be noted that the curves for two exemplar students were not only the product of the author's imagination but have been identifiable in real educational data. For instance, Student A grew slower in the initial stage but caught up in a faster rate in the later timepoints, which can be observed in learners with special education needs or English language learners. Student B's ability grew almost linearly, which has been observed in the medical professions education data (Dumas, McNeish, Schreiber-Gregory & Durning, 2019). When the overall model fit indices of the inverted-J function are acceptable, every student (including student A and B) then receives an estimate of asymptotic capacity. Nevertheless, the asymptotes for both cases are suspect because neither growth curve has leveled off toward the capacity within the available timepoints. In this case, interpretations of their capacity are not appropriate, and researchers should try to avoid any misinterpretations about how much students such as these can learn in the future.

Within the present DMM framework, despite that the random effect term in DMM could capture differences between the student-specific growth parameter and the population-averaged parameter estimate, it does not indicate the appropriateness of DMM as a measurement model to gauge learning capacity for each student. No statistical index has been developed to evaluate the model appropriateness for each student as well as the aberrance of student growth from a typical growth trajectory. There is a need for a person-specific fit statistic to help researchers identify students with aberrant growth and thus avoid misinterpretation of the student-specific DMM parameters. Further, achieving high reliability of learner-specific capacity estimates could be challenging due to existing limitations in a dataset (e.g., only a few available data points, too much missing data, or very small intervals between timepoints), and removing deviant cases may increase the reliability of the DMM scores as well as enhance other validity evidence.

Research Goals of the Current Project

The current project consists of two parts. Given the theoretical commonalities between DMM and IRT models (e.g., both can be conceptualized as scoring models), this research reviewed and adapted existing formulation of IRT person-fit or model appropriateness statistics into the DMM framework. The study first formulated four versions of such a statistic within the DMM framework (i.e., Trajectory Deviance Index, TDI). To determine the final version of the TDI among four candidates, their distributional characteristics (e.g., mean, standard deviation, and form of distribution), empirical critical values, and effectiveness (e.g., power, Type-I error rate) were examined and compared via simulation investigations.

The second part of the project studied the influence of the *Building Blocks* intervention as well as student characteristics (e.g., special education status, bilingual status, SES, gender, and age) on the deviance of growth in early mathematics by applying the developed TDI to the TRIAD data. Each TRIAD student received an estimate of TDI, and misfitting students were identified based on the empirical critical values at the fifth percentile in the previous simulation. The study further examined and discussed whether incorporating the TDI improved the DMM conditional reliability and other aspects of validity.

Contribution to the Field

Dynamic Measurement Modeling (DMM) enables educational researchers to gauge learning capacity, which could be critical to improving current measurement practices. The developed index provides researchers with an effective way to check the individual appropriateness of DMM models. It helps researchers to identify the students who do not fit the DMM and therefore avoid misinterpretation of the estimated parameters for these students. Removing the students with a deviant trajectory from DMM can also improve the reliability of the estimates for other students and further enhance the validity of DMM, although it is an optional step for using DMM. The TDI is expected to be a useful statistical tool for substantial research in the future. For example, researchers can explore influential factors (e.g., demographics, intervention conditions) that contribute to the TDI across different constructs, domains, and measurement contexts. In other words, what makes students deviate from a typical growth trajectory: a critical issue related to the general mission of social-justice and testing fairness of DMM research and the larger educational research field.

Chapter Two: Literature Review

This chapter synthesizes the literature from three areas. It starts with a review of related issues of achievement gaps and how previous measurement practices contributed to the persistence of different achievement gaps. The second section introduces the DMM methodology and elaborates on how DMM may help reduce the gaps based on the available literature. Finally, after reviewing the previous person-fit or model appropriateness statistics research from other measurement frameworks (e.g., IRT), four DMM trajectory deviance indices are conceptualized and formulated.

Achievement Gaps and Conventional Measurement Practices

Both policy and research efforts have been made to close the U.S. achievement gaps for decades, but the gaps remain apparent (Reardon, 2011; Scammacca et al., 2020). Since the beginning of the 21st century, initiatives such as the *No Child Left Behind*, the *Race to the Top,* and the *Every Student Succeeds Act* were enacted with the stated purpose to support low achievers and narrow down achievement gaps. Unfortunately, according to an investigation of the National Assessment of Educational Progress (NAEP, Education Commission of the States, 2018), the average gap between African American and Hispanic students and their White counterparts was substantive, and the gap was even larger among low performers. The NAEP results also indicated that the gender gap had not substantially narrowed since the first assessment back in 1994. Researchers who have investigated the underlying causes or "input" of achievement gaps have offered a variety of explanations and names for the phenomenon of achievement gaps. For example, given the disparities in educational opportunities, Flores (2007) reframed the issue of achievement gaps into the Opportunity Gaps. Chambers (2009) used the term, Receivement Gap, to emphasize the input instead of the output of the gaps.

Different Types of Achievement Gaps

In the context of education, the achievement gap issue is often specified into several categories from the results of comparing different subgroups of students. We primarily discuss three types of achievement gaps, including a racial gap, socioeconomic status (SES) gap, and gender gap. These gaps are not independent of each other, and their interaction effects were also commonly investigated in research practices (e.g., Harackiewicz et al., 2016; Quinn & Cooc, 2015).

Race/Ethnicity Gap. The racial gap often refers to the disparities in which African American and Hispanic students had lower performance than their White counterparts on standardized tests within the United States (Milner, 2013). This type of achievement gap was not only observed at the entry to school but also found to shift (either widening or narrowing) across school years (Fryer & Levitt, 2006; Henry et al., 2020). The academic achievement of Asian American students has been observed to be close or slightly higher than that of White students (Education Commission of the States, 2018; Kao & Thompson, 2003), so they are not perceived as a historically disadvantaged minority in the same way as Black and Hispanic students in the research of racial achievement gaps in the U.S. (e.g., Potter & Morris, 2017). Researchers have been exploring reasons that attributed to the issue, and theories or sources were formed. Three contributing factors are briefly reviewed here: stereotype threat; opportunity gaps; and family experiences.

Stereotype Threat. Stereotype threat refers to a phenomenon that an individual's social identity (e.g., race, gender) is devalued in a given setting (Steele, Spencer & Aronson, 2002). It has been recognized as a contributing factor to the persistent racial/ethnic achievement gap (Osborne, 2001; Steele, 1997). Specifically, African American and Hispanic students usually contend with the stereotypes that students from their ethnic group are less likely to succeed in academic settings than their counterparts (Aronson, 2002; Steele & Aronson, 1995). Such threats have longitudinal impacts on these students (Steele et al., 2002). Students may feel chronically stressed, anxious, and distracted (Schmader et al., 2008), and all these psychological feelings and responses negatively influence the academic performances of the students who were stereotyped (Cohen & Garcia, 2008).

Opportunity Gap. Disparities in educational opportunities have been perceived as the underlying causes of the racial gap (Flores, 2007; Kuhfeld, Gershoff & Paschall, 2018). This perspective emphasized that learning opportunities (or inputs) were not equally distributed among students from different racial/ethnic groups. For example, African-American and Hispanic students were less likely to access qualified and experienced teachers but more likely to receive low expectations for academic achievement (Flores, 2007; Wilkins et al., 2006). The importance of teacher qualification and expectations of students' learning gains (or outputs) have been supported and explained by a variety of theories (e.g., Pygmalion effect in classrooms, Rosenthal, & Jacobson, 1968) and research (e.g., Phillips, 2010; Telese, 2012). The inequitable distribution of inputs were found to have resulted in the observed gaps in the outputs across racial groups.

Family Experiences. Another reason for the racial gap was racial disparities in various aspects of family experiences such as parenting style (e.g., notion of concerted cultivation, Lareau, 2003/2011), parental involvement (e.g., Cheadle & Amato, 2011), parental expectations (e.g., Davis-Kean, 2005; Davis-Kean & Sexton, 2009), family investment in education (e.g., Cheadle, 2008), and so forth. Potter and his colleagues (Potter & Roksa, 2013; Potter & Morris, 2017) found that family experience tended to be stable over time, and therefore the impact of family experiences was cumulative rather than a one-time occurrence. As a result, the consistent differences in family experiences may exacerbate the achievement gap among racial groups (DiPrete & Eirich, 2006; Potter & Morris, 2017). A body of research has shown that Black or Hispanic families were in a disadvantaged position compared to their White or Asian counterparts (e.g., Cheadle & Amato, 2011; Pattillo-McCoy, 1999), although the findings of racial disparities in family experiences are mixed in general.

Moreover, the issue of racial disparities in family experiences has interacted with SES (e.g., Henry et al., 2020; Kuhfeld et al., 2018), so both race/ethnicity and SES issues may be involved in explaining how family experiences are attributed to achievement gaps. For example, Lareau (2003) termed *concerted cultivation* as a social-class related parenting style, which subsumes parental school involvement, students' enrollment in extracurricular activities, and the amount of educationally beneficial materials in the home. Specifically, parents from higher SES classes are more likely to engage in concerted cultivation while lower-class parents often engaged in the "accomplishment of natural growth", and such differences in parenting style can reproduce social class intergenerationally. In the U.S., racial groups have been historically closely related to social class, and recent data shows that more students from Black and Hispanic groups were living in poverty than White and Asian counterparts (U.S. Census Bureau, 2018). Given that SES accounts for a substantial amount of variance in educational achievement (Duncan & Magnuson, 2005; Magnuson & Duncan, 2006; von Stumm, 2017), the disparities in parenting style have contributed to the sustainment of both racial and SES achievement gaps. Even with SES controlled, racial background was still strongly related to the levels of concerted cultivation that parents engaged in (Cheadle, 2008; Cheadle & Amato, 2011), which indicates that race/ethnic gaps in advantaged parenting style were not merely manifestations of SES.

SES Gap. The SES gap refers to the disparities in academic achievement among students with different SES backgrounds, that is, students from higher SES families perform better on achievement tests, and it even appears to have larger influences on academic achievement than racial and ethnic backgrounds (Hadden et al., 2020). SES is either measured through a combination of indicators such as highest levels of parents' education, occupation, family income, poverty status, household books, and so forth (Dumas & McNeish, 2017) or represented by a single indicator (e.g., occupation, free/reduced-price lunch) in research settings (Duncan & Magnuson, 2005). As pointed out by Duncan and Magnuson (2005), SES was an umbrella term because it not only refers to an individual or a family's relative position in a social hierarchy but also their privileges to access a variety of socioeconomic resources.

In a recent investigation, Chmielewski (2019) studied the SES gap using the longitudinal data of 30 large-scale assessments with 5.8 million students across 100 countries. It was concluded that the SES achievement gap increased globally over the past 50 years (1964-2015), and empirical evidence was found to support a declining trend of the gap in the United States. Given that the SES gap has been shown to increase in previous studies (e.g., Reardon, 2011), the new evidence can be perceived as a good signal or a credit for the efforts that educators have made to close achievement gaps and address educational equity. Nevertheless, the declining trend does not indicate that the disparities in achievement between students from low- and high-SES groups have been eliminated already. The magnitude of the current SES gap is undoubtedly still concerning.

Gender Gap. Unlike the racial- and SES- gaps, the gender gap shows a mixed pattern across domains. While female students have outperformed male students in the domain of literacy (Organization for Economic Cooperation and Development, OECD, 2015; Rvachew et al., 2020), they are relatively under-represented in the fields of science, technology, engineering, and mathematics (STEM, Hill, Corbett & Rose, 2010; Halpern, 2014; Jungert et al., 2019; Robinson-Cimpian et al., 2014). Explanations for this longstanding issue have been complex and shifted over time.

Cognitive psychologists detected that girls performed significantly better than boys on measures of executive function (e.g., inhibitory control, Altemeier et al., 2008; Carlson & Moses, 2001). Given that executive function (EF) fundamentally supports the development of literacy skills (Best et al., 2011; Ribner et al., 2017), the observed EF differences were claimed to be the reason for the gender gap in literacy. However, this explanation may not be tenable when taking the research about EF and early math learning into account (Rvachew et al., 2020). The positive correlations between EF and early mathematics skills have been well-established (Clements et al., 2016; Dong et al., 2020). As mentioned, another aspect of the gender gap is that male students outperformed female students in mathematics and other math-intensive domains. Therefore, using EF as a major explanation of the gender gap in literacy achievement was inconsistent between domains.

In the late 20th century, researchers were still claiming that males had higher talent or aptitude in mathematics and science (e.g., Benbow ,1988; Nowell & Hedges, 1998). However, a variety of new evidence has demonstrated that this kind of claim cannot be supported. For example, the genetic or biological foundation of the core systems for mathematical and scientific thinking are equally available to males and females (Spelke, 2003; Spelke, 2005). Also, there was no substantial gender difference in learning capacity among the genders observed via DMM (Dumas & McNeish, 2017).

Instead, discrepancies in learning opportunities, teacher perceptions, and other contextual factors (e.g., stereotyping) were more valued in explaining the gender gap. The gender-roles women historically played have been influences on their expectations for success, career choice, and learning opportunities in the fields of math and science (Eccles, 1986). Moreover, teachers' biased perceptions such as "boys are less competent in language arts than girls" or "boys have better mathematics skills than girls" may have increased the gender gap (Robinson-Cimpian et al., 2014). Those gender stereotypes from teachers could be transmitted to students (e.g., gender ability stereotype endorsement, Plante et al., 2019). Once learning became gender-related in the minds of some parents and educators, the gender gap was then naturally sustained.

Closing achievement gaps is an ongoing and challenging task for researchers and the entire education community. Besides the aforementioned types of gaps, disparities in academic achievement also occur between mainstream students and other minorities (e.g., English language learners, students with disabilities). Although they were not delineated in detail here, the study does not aim to discount the importance of these issues in the context of education.

How do Educational Testing Practices Influence Achievement Gaps?

The achievement gaps are usually scrutinized via different formats of achievement data (e.g., test scores, course grade, and graduation rate) collected from single-timepoint tests. The validity of measurement tools for collecting achievement data has been a fundamental issue in the research of educational testing (AERA, APA, & NCME, 2014), and lack of validity can lead to a variety of negative consequences. Additionally, test-related educational policy (e.g., No Child Left Behind, NCLB) also played a role in the persistence of the gaps. Both validity- and policy-related testing issues that may have contributed to the appearance or changes of achievement gaps are briefly reviewed here.

Measurement Invariance of Educational Assessment. Measurement invariance (MI) generally indicates that the underlying constructs are measured in the same way across groups or measurement occasions (Meade et al., 2008; Meade & Lautenschlager, 2004), which has been perceived as an important assumption required to validly compare participants' scores across groups (Dong & Dumas, 2020). An educational assessment that fails to demonstrate MI could generate biased scores for students from a subgroup. For example, a "Bus Pass" problem was presented to a group of African American middle school students in a districtwide math test (Tate, 1994):

"It costs $ 1.50 each way to ride the bus between home and work. A weekly pass is 16.00. Which is the better deal, paying the daily fare or buying the weekly?"

One of the key factors to answer this question correctly was having the notion about how many days people work per week. The test developer and students from typical middle-class families are more likely to reach the same notion: people work 5 days per week. However, for those growing up in lower SES families, their life experience (e.g., parents' working schedule) could inform them that people work 6 or even 7 days a week. Therefore, middle-class students are more likely to give the "correct" answer that was provided by the test developer, which means this math problem did not function in the same way for students with different SES backgrounds. Utilizing such measures or items can mislead teachers' evaluation of students' math competencies and reduce the accuracy of detecting achievement gaps in research practices.

In addition to the content-based bias, the format of tests was also demonstrated to be one reason for the failure of establishing MI. Research shows that item-format was associated with the performance of students from different gender groups in both math and reading tasks (Reardon et al., 2018; Schwabe et al., 2015). For example, male students favored multiple-choice items, while girls performed better on constructed-response items. Reardon et al. (2018) pointed out that the test format accounted for about 25% of the variance in gender achievement gaps.

Test Accommodations. MI is not always able to be established over different populations, so test accommodations have been frequently used as an alternative way to improve test fairness and validity (e.g., Abedi et al., 2004; Li & Suen, 2012). English language learners (ELLs) constitute a large student population in the United States, but many developed standardized tests that ELLs took later were originally developed for native English speakers (Abedi & Gándara, 2007). If a measure aims to gauge students' mathematics achievement, ELLs with insufficient English proficiencies are likely to misunderstand math problems written in English. In this case, the underlying construct measured by the test was English proficiency rather than mathematical competency, and measurement invariance and other validity aspects of this measure cannot be guaranteed. Consequently, the achievement gap between ELLs and mainstream students can be over-estimated. This issue cannot be easily solved by translation (Hambleton, 2005; Solano-Flores, 2011), and therefore there is a vein of research focusing on different test accommodations for ELLs. A meta-analysis of the effects of test accommodations for ELLs revealed that test accommodations were generally effective to improve ELLs' performance, especially for those with a low level of English proficiency (Li & Suen, 2012). Given that effective accommodations are capable of improving the accuracy of evaluating students' performance as well as recognizing achievement gaps, they can be perceived as a remedy to threats to measurement validity.

Teaching to the Test. In the NCLB[1] era, policies, curriculums, and instruction shifted with a focus on standardized testing, and "teaching to the test" became a daily practice in many schools and classrooms across the nation (Davis & Martin, 2008; Menken, 2006). Because of the pressure that teachers received about improving students' scores on high-stakes assessments, they were driven to provide test-centered content and instruction to students (e.g., item-teaching, Popham, 2001; test-taking skills, Jennings & Bearak, 2014). As a consequence, students gained more in test scores than learning in the domain. With the longstanding goal of narrowing achievement gaps, such policies pushed the disadvantaged groups (e.g., low SES, African American, and ELLs) to a worse position because they were more likely to receive the "teaching to the test" instruction. For example, Davis and Martin (2008) shared the fact that district and school administrators in Baltimore implemented supplemental math programs which primarily aimed to prepare African American students for the state-wide tests. Menken (2006) also found that some high schools in New York had replaced native language instruction with test preparation strategies for ELLs. Despite that the NCLB act was switched to the Every Student Succeeds act in 2015, the sequelae have not yet vanished.

[1] *No Child Left Behind Act,* Public Law 107-110 (2001)

In summary, testing practices should have played a more positive role in promoting test fairness and closing achievement gaps. Reformation in the current system of educational measurement with effective approaches, criteria, and guidelines is needed. The *Standards for Educational and Psychological Testing* (also called Standards, AERA, APA, & NCME, 1999, 2014) aims to provide a basis for evaluating the quality and appropriateness of testing practices, including test development, test use, interpretation and so forth. In the latest 2014 version of *Standards*, testing fairness is now listed as a fundamental issue as to validity and reliability in the measurement research (Plake & Wise, 2014). To align with this scope, generating less biased estimates should be kept as a primary goal of future measurement work.

Employing a Dynamic Measurement Modeling Framework in Educational Testing Practices

In educational measurement practices, conventional single-timepoint tests are usually applied to gauge developed constructs (e.g., cognitive abilities, Sternberg et al., 2002), but these test scores only indicate the construct levels prior to or at the time of test administration. Educators and researchers always believe in students' potential to learn and grow (Dweck, 2015; Vygotsky, 1931/1997; Vygotsky, 1978), but critical educational decisions (e.g., school funding, college admission) are usually made according to their developed abilities (i.e., how much they have learned). As documented by a fair amount of studies regarding achievement gaps (e.g., Flores, 2007; Kirsch et al.,2007; Reardon, 2011; Quinn, 2015), students from certain demographic groups or growing up in disadvantaged circumstances showed less developed abilities measured by single-timepoint tests. The achievement gap is sustained because students with low performance in these tests received fewer learning resources and opportunities. Therefore, in order to improve measurement practices as well as make equitable decisions, educational researchers have shown explicit interest in seeking effective indicators or constructs other than the developed abilities.

One remarkable construct is *Learning Capacity*, which refers to the maximum amount of ability that an individual can be expected to attain in a specific domain (e.g., elementary and middle-school mathematics, Dumas & McNeish, 2017; McNeish & Dumas, 2017; McNeish et al., 2019). In contrast to those developed constructs, learning capacity is a developing construct that reflects how much students can learn in the future from the time of assessment administration (Sternberg et al., 2002). Dumas and McNeish (2017) developed a formal statistical modeling framework, *Dynamic Measurement Modeling*, to measure learning capacity by producing learner-specific asymptotes. Meanwhile, they also empirically demonstrated that those capacity scores from DMM were much less impacted by students' socioeconomic status, race/ethnicity, and gender than conventional achievement scores, which indicates that employing DMM in the educational system could be a critical step to improve the consequential validity of current educational measurement (Dumas & McNeish, 2017, 2018). In this section, previous work regarding the DMM as well as the historical perspectives of dynamic measurement are reviewed.

Development of Dynamic Measurement Modeling

How are we to assess learning capacity and other developing constructs? This question has concerned psychologists for a long time. At the beginning of the 20th century, Alfred Binet advocated measuring students' capacity for performance (e.g., a process assessment paradigm), but his major contribution to psychological measurement was still about the single-timepoint testing work for gauging developed constructs (also termed *static* tests, Dumas, McNeish & Greene, 2020; Sternberg & Grigorenko, 2002). For example, Binet's most well-known contribution, the Binet–Simon intelligence test, was created to measure a developed or current intelligence level for examinees. However, the developing feature of the intelligence construct has also been widely recognized (e.g., the theory of fluid and crystallized intelligence, Cattell, 1963; Piaget's theory of cognitive development, Hooper, Fitzgerald & Papalia, 1971), and formal instruction, experience, and educational context are all important to the development of intelligence (Horn, 1967). Early psychologists had already reached the conclusion that conventional static tests (e.g., the Binet–Simon intelligence test) could only measure the developed abilities rather than the associated developing constructs (i.e., capacities of cognitive abilities, Rey, 1934; Vygotsky 1934/1962).

Researchers may argue that examinees' current performance in static tests is often a strong predictor of their future performance, and therefore it is not necessary to measure the developing constructs via advanced approaches. However, such a claim has been criticized. As early as 1920, DuBois (1920/2013) highlighted the importance of incorporating information regarding learning or growing processes into large-scale testing practices because the static tests would always benefit students who entered with higher performance ("rich get richer"). DuBois' argument was strongly supported by later studies, as shown in the literature of achievement gaps previously.

The Dynamic Assessment Conceptual Model. Given that traditional static or single-timepoint tests cannot meaningfully capture learning capacity and other developing constructs (Rey, 1934; Vygotsky 1934/1962), researchers have explored alternative testing practices and methodologies to achieve the goal. However, it was not well addressed until Israeli cognitive psychologist Feuerstein (1979) established the *Dynamic Assessment* (DA) framework. DA is capable of studying student learning capacity through multiple testing occasions with standardized instructions integrated (Feuerstein, 1979; Tzuriel, 2001). Notably, the term *dynamic* was then commonly used for the later measurement work featuring non-static testing practices (e.g., dynamic measurement modeling, Dumas & McNeish, 2017; dynamic testing, Sternberg & Grigorenko, 2002) Dumas et al. (2020) advocated for use of this term to refer to measurement practices aimed at developing constructs through multiple-timepoint testing and meanwhile considering the instructional and environmental contextual factors that support the developmental process.

The idea of DA originally derived from Feuerstein's work on sorting child survivors of Nazi concentration camps into appropriate grade levels immediately after World War II (Feuerstein, Feuerstein, Falik & Rand, 2002; Feuerstein Krasilowsky & Rand, 1974). They found that single-timepoint static measures underestimated the learning potential of those child survivors. Importantly, it was believed that their experience in concentration camps might have delayed their current abilities but not necessarily their learning capacity. Feuerstein then raised the idea of testing cognitive abilities through multiple-timepoint tests with learning opportunities and high-quality instruction integrated. This paradigm strongly aligns with the contemporary vision of filling the opportunity gaps for the low achieving students from disadvantaged groups (e.g., Flores, 2007).

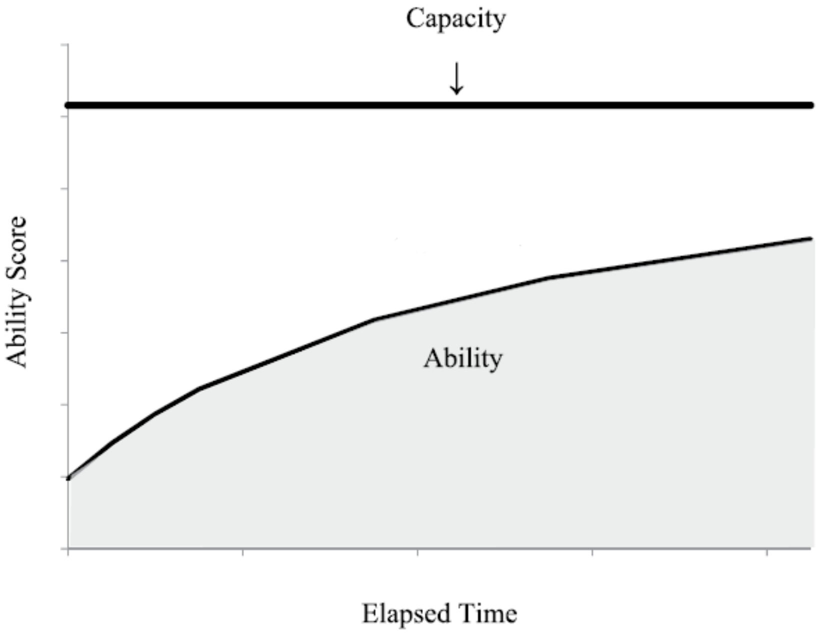


Figure 2. Theoretical Depiction of the Process of Dynamic Assessment

After mapping out the growth trajectories of students across timepoints, Feuerstein (1979) observed that student ability grew nonlinearly and flattened out as time passes. The theoretical diagram of DA is displayed in Figure 2. As can be seen, the nonlinear line indicates the growing ability for each examinee over time, and the space below the curve and the horizontal line can be conceptualized as the developed ability. Meanwhile, the horizontal line at the top is the potential or capacity, which is a developing construct. Notably, the capacity in the DA framework is conceptually relevant to but fundamentally different from *aptitude* (Lohman, 1999; Thompson & Zamboanga, 2004. As stated by Dumas and McNeish (2017), aptitude is usually considered as existing in students without implementing specific instructions, while capacity is closely related to a student's environment and educational experience, and only shows after a series of repeated testing with integrated instructions.

Despite that fact that the initial conceptual model was established, there were several issues that hindered the application of DA in large-scale research or practical work in the fields of education and psychology. First, the operational costs of DA methods are typically higher than static tests. Specifically, estimating capacity through DA requires multiple testing occasions, and collecting longitudinal data for each participant is more expensive and less feasible than data collection for conventional static tests. Moreover, the integrated instructional interventions among timepoints may cost additional resources for clinical training and professional development. Second, Feuerstein's DA theory was generally perceived as a fundamental conceptual framework, although he had applied it in his research work (Feuerstein et al., 1981; Feuerstein et al., 1987). Later researchers encountered both statistical and psychometric challenges in quantifying the DA components (e.g., capacity in Figure 2, Embertson, 1987; Sijtsma, 1993). As indicated by the quote at the beginning of this study (Sternberg & Grigorenko, 2002), educational researchers by that time were still looking for a formal methodology framework to realize those important concepts from DA.

Dynamic Measurement Modeling. To fill this research gap, Dumas and McNeish (2017) created DMM to measure learning capacity from developed ability scores at multiple testing occasions. DMM utilizes a nonlinear mixed-effects modeling framework that formalizes DA concepts into a statistical model and generates learner-specific upper asymptotes for quantifying capacity. Within the DMM framework, it does not necessarily require the standardized but resource-intensive instructions between test occasions as in DA. Instead, the DMM paradigm capitalizes on existing practices such as classroom instructions, research interventions, and other educational programs that may influence student growth (Dumas & McNeish, 2018; Dumas et al., 2020). DMM is thus capable of producing learner-specific capacity asymptotes within large-scale educational data, including previously collected secondary data for other research purposes.

Contributions of DMM to Research and Testing Practices**.** As previously mentioned, one major contribution of DMM is to improve the consequential validity of assessments through producing estimates of learning capacity and other growth parameters (e.g., learning rate). Specifically, Dumas and McNeish (2017, 2018) applied DMM techniques with the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) data of the 1998-1999 cohort, for both mathematics and reading assessments. The estimated capacity scores from DMM were much less influenced by a combination of demographic factors (e.g., socioeconomic status, race/ethnicity, and gender) that were usually associated with achievement scores from conventional static tests. Notably, dynamic testing has shown to be superior in capturing developing constructs than static testing, but they are not incompatible. The quality of each single-timepoint test may impact the success of using dynamic testing to measure students' learning capacity. These two sorts of testing practices (static and dynamic) are "on a continuum" in effect (Sternberg & Grigorenko, 2002). It is rare to see dynamic assessment researchers discard the importance of static testing, instead, they conceptualized capacity from DMM as a higher-order or meta-construct for the ability scores measured at each static testing occasion (e.g., McNeish & Dumas, 2018) .

The application of the DMM method is beyond a tool for producing estimates of learning capacity, and this newly innovated methodological framework has been used to investigate a wide range of substantive problems across domains. In the study of early mathematical learning (Dumas, McNeish, Sarama & Clements, 2019), DMM showed the capability of taking both intra- and inter-individual differences in student growth trajectory into account. It was found that the *Building Blocks* (BB) intervention significantly improved students' learning rate, especially for minority students (e.g., Black or Latinx). Moreover, there was no statistical difference in the asymptotic capacity estimates across intervention conditions, which implies that students' learning capacity was not impacted by received learning opportunities and resources (e.g., high-quality intervention in this case). Therefore the capacity estimates may serve as a unique piece of evidence for evaluating student performance. McNeish et al. (2019) also demonstrated that DMM capacity scores (43% variance explained) outperformed conventional ability scores (16% variance explained) estimated by traditional IRT models in the prediction of adult verbal scores, which further verifies the effectiveness of using DMM-based capacity estimates to forecast students' future performance.

DMM has also contributed to investigations of summer learning loss, a critical but still under-researched topic in the U.S. Given that there were limited statistical methods for estimating learning loss, McNeish and Dumas (2020) recently incorporated and quantified summer learning loss through their DMM model. This work is anticipated to help researchers meaningfully capture and interpret summer learning loss. Additionally, the DMM paradigm was successfully applied to the field of medical professions education and training (Dumas, McNeish, Schreiber-Gregory & Durning, 2019).

Specifying and Fitting a DMM

To specify and fit a DMM, the empirical example used in this project was the Technology-enhanced, Research-based, Instruction, Assessment, and professional Development (TRIAD, Sarama & Clements, 2013; Sarama et al., 2008) cluster-randomized experiment data, which contained seven waves of mathematics test scores: pre and post of PreK, Kindergarten, 1st grade, 3rd grade, 4th grade and 5th grade. So, while the procedure reviewed here can be generalized to any suitable dataset, this study paid special attention to issues related to fitting DMM to the TRIAD data.

Shape of Growth. A functional form or shape of growth needs to be selected to fit a DMM to any dataset. It should be noticed that the nonlinear function of DMM can be specified and parameterized in a variety of ways in order to address specific research questions (Dumas & McNeish, 2017), which may result in different shapes of growth trajectories. However, the shapes should come with interpretable upper asymptotes, and thus nonlinear growth models with “S-shaped” trajectories (e.g., logistic) or “inverted J-shaped” (e.g., exponential, Michaelis-Menten) are often considered for DMMs (McNeish

& Dumas, 2017). The Michaelis-Menten curve (Michaelis &Menten, 1913), a well-known model of enzyme kinetics in biochemistry, posited an inverted J-shape growth trajectory, in which student ability grows in a nonlinear way and eventually reaches an upper asymptote (learning capacity) over time. This growth model theoretically matched with the DA based conceptualization of ability growth and also empirically fitted to the TRIAD data (Dumas, McNeish, Sarama & Clements, 2019). In the initial DMM work conducted by Dumas and McNeish (Dumas & McNeish, 2017; McNeish & Dumas, 2017), it was also shown to be the best-fitting functional form to ECLS-K mathematical data among several competing models such as logistic, Gompertz, Richards, and von Bertalanffy.

Parameterization. In the Michaelis-Menten function, for the *i*th student in the TRIAD data set (*i* = 1, . . ., *N*) at the *t*th timepoint (*t* = 1, . . ., *T*) can be written as:

$$Math_{ti} = \beta_{0i} + \frac{(\beta_{Ai} - \beta_{0i})Time_t}{\beta_{Ri} + Time_t} + d_{ti}\,, \qquad (1)$$

where

$$\begin{aligned} \beta_{0\mathrm{i}} &= \alpha_0 + \zeta_{0\mathrm{i}} \\ \beta_{R\mathrm{i}} &= \alpha_R + \zeta_{R\mathrm{i}} \\ \beta_{A\mathrm{i}} &= \alpha_A + \zeta_{A\mathrm{i}} \end{aligned} \qquad (2)$$

As can been seen in Equation (1), three parameters of the Michaelis-Menten model are: the initial value (β_{0i}), which captures initial ability when time is 0 (e.g., a pre-test of a longitudinal dataset); the learning rate parameter (β_{Ri}), which captures the point in time when ability is halfway between the initial value and the upper asymptote; the learning capacity parameter or upper asymptote (β_{Ai}), which captures the maximum ability growth as time approaches infinity. d_{ti} is a residual term that represents the difference between model-predicted and observed values at each time point for each student. From Equation (2), each of the DMM parameter β consists of a population-averaged fixed effect term (α) and a student-specific random effect term (ζ_i) that allows each student to have their unique growth curve. Figure 3 depicts DMM parameters and presents the relations among them.

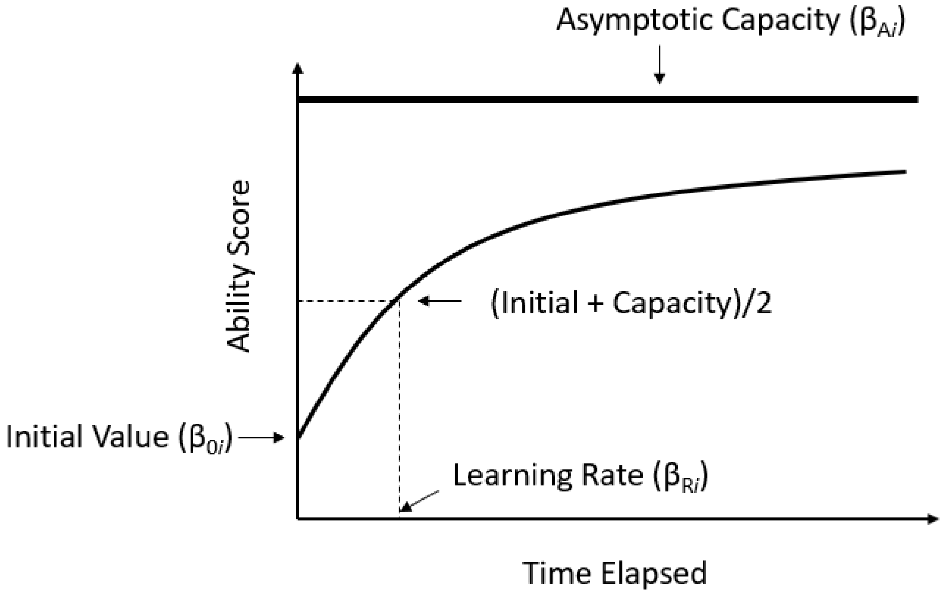

Asymptotic Capacity (β_{Ai})
Ability Score
(Initial + Capacity)/2
Initial Value (β_{0i})
Learning Rate (β_{Ri})
Time Elapsed

Figure 3. A Hypothetical Learner-Specific DMM Curve

Reliability for DMM Capacity Estimates. McNeish and Dumas (2018) has advanced procedures in computing reliability for DMM capacity estimates. DMM is substantively different from either classical test theory (CTT) or IRT frameworks, but it still shares theoretical aspects in common with typical measurement models. For example, DMM can be conceptualized as a scoring model, in which capacities are estimated from those longitudinal ability scores (McNeish & Dumas, 2018). While a measurement model employs observed indicators to measure a latent construct, DMM utilizes vertically scaled scores (*Math*$_{ti}$) to estimate learning capacity. Analogically, the ability scores in DMM were treated as observed variables without measurement error (McNeish & Dumas, 2018). After evaluating a series of methods for generating a CTT or IRT reliability index, McNeish and Dumas adapted a conditional reliability function outlined in Nicewander (2018) to DMM by replacing the appropriate information function:

$$\rho_{XX'} \mid \beta_{Ai} = 1 - \frac{Var(\hat{\beta}_{Ai})}{Var(\zeta_{Ai})} \qquad (3)$$

In Equation (3), $Var(\hat{\beta}_{Ai})$ is the variability for the *i*th student's capacity estimates in the TRIAD dataset, and $Var(\zeta_{Ai})$ is the variance of all capacity random effects. As can be seen, the function relies on the asymptote parameter quantities, meaning that the utilization of the DMM conditional reliability index is not restricted to a specific growth trajectory (e.g., the Michaelis-Menten curve for the TRIAD example), but can be used for any functional form with an upper asymptote.

Developing a Trajectory Deviance Index for DMM

DMM has shown its capability of capturing developing constructs and demonstrated better consequential validity than static testing (Dumas & McNeish, 2017; Dumas & McNeish, 2018), but the refinement of this newly invented methodological framework is a continuing task. The accuracy of estimates is essential to all measurement models, especially for those used in educational decision-making (Albers et al., 2016; Karabatsos, 2003; Meijer & Sijtsma, 2001). Although the extant DMM method can produce reliable and valid estimates of learning capacity across domains (e.g., Dumas, McNeish, Sarama & Clements, 2019; McNeish, Dumas & Grimm, 2019), there is room to further improve the reliability and accuracy of estimates, such as incorporating a person-specific fit statistic to identify students whose growth trajectories do not align with a model-implied growth function.

This section maps an analogical relation between IRT and DMM in order to show that existing person-fit indices from the IRT framework may be able to be meaningfully adapted into DMM in order to understand which student deviate from the overall trajrctory of learning. Here, I seek candidate indices that may be adapted into the DMM framework. CTT procedures for determining person fit also exist, but they are not delineated in detail for three specific reasons. First, IRT (also termed latent methods here) has been generally considered to be superior to CTT. DMM is also a latent variable method (Dumas & McNeish, 2017), and therefore the functional components in DMM are comparable to those in the IRT framework. Second, person-fit in CTT depends on the sample of examinees, or requires normative comparisons, which largely limits its application in different measurement settings. Third, person-fit methods building on IRT have been better-established than CTT methods (see Karabatsos, 2003; Meijer & Sijtsma, 2001). Given the richness of person-fit research within the IRT framework, it has been common for methodologists to extend those indices for use in other contexts or frameworks. For example, l_z (Drasgow, Levine & Williams, 1985), a well-known likelihood-based IRT model appropriateness index, were adapted for cognitive diagnostic models (Cui & Li, 2015; Santos et al., 2020).

Conceptualizations of Person-Fit between IRT and DMM

A basic item-response (or item-score) pattern in IRT is that, given a student (*i*) answering a set of test items, there are higher probabilities for the student to correctly answer easier items and lower probabilities for the student to correctly answer more difficult items. Person-fit or model appropriateness indices in IRT were designed to evaluate the misfit of a student's testing performance to item-response patterns and to quantify the reasonableness of a student's answers to items within a test (Karabatsos, 2003). As specified in Equation (1), although the DMM method reflects important features of a growth model, it can also be conceptualized as a scoring model (i.e., a longitudinal psychometric model, McNeish & Dumas, 2018). Then, statistics for DMM that analogically resemble person-fit can be built on this essential theoretical conceptualization.

Several key components of the two types of models are comparable and share similarities. Both are described as measurement models to capture latent constructs. Although IRT models often use test items as indicators to assess examinees' abilities, DMM utilizes examinees' ability scores across timepoints to capture their learning capacity in the corresponding ability domain. Compared to an IRT-based index, an analogical DMM person-fit statistic must therefore evaluate the model appropriateness for each examinee, but it should focus on detecting the aberrance of student growth from a model-implied growth pattern (e.g., Michaelis-Menten trajectory) rather than an item-response pattern. The term “person-fit” is frequently used by Rasch researchers (e.g., Wright & Stone, 1979; Wright & Masters, 1982; Linacre, 2020), whereas other IRT researchers studying 2 or 3-parameter logistic (PL) models usually prefer to term such a statistic as “appropriateness index” (e.g., Drasgow, Levine & Williams, 1985). This study uses two terms interchangeably. In contrast, given the conceptualization of person-fit within the DMM framework above, a DMM person-fit statistic is termed as a Trajectory Deviance Index (TDI).

TDI for DMM: Implications from IRT Framework

Since the widespread use of IRT models in the late 1970s, many person-fit statistics have been developed to detect misfit of item-score patterns. Meijer and Sijtsma (2001) reviewed available person-fit methodologies in a systematic way and summarized 23 different person-fit methods into four general categories: Rasch, 2 or 3-parameter logistic (PL) models, computerized adaptive testing (CAT), and group-based/nonparametric. Later studies identified some additional person-fit statistics and conducted a slightly different categorization (e.g., parametric vs. non-parametric, Karabatsos, 2003; Rupp, 2013), but the difference in categorization of the statistics barely impacts the meaning or utilization of those methods. Given that multiple researchers have done delicate work on reviewing the available person-fit indices (see Appendices in Karabatsos, 2003 or Meijer & Sijtsma, 2001), the present study does not delineate the formulas, notations, and technical details of every method here. Instead, representative examples are described in the following paragraphs to help illustrate specific problems.

As indicated in the person-fit literature, the fundamental methodology research were mostly accomplished in the 20th century (e.g., l_0, Levine & Rubin, 1979; l_z, Drasgow, Levine & Williams, 1985; *U*, Wright & Stone, 1979; *W*, Wright & Masters, 1982; Z_C, McLeod & Lewis, 1999). In recent years, new person-fit statistics were mostly about adaption, transformation, or adjustment of the previous methods to address specific research needs. For example, Marianti et al. (2014) adapted the IRT-based l_z (Drasgow, Levine & Williams, 1985) into l_z^t for response-time models, and Sinharay (2018) later developed a new person-fit index by replacing the τ component in l_z^t by its maximum likelihood estimation ($\hat{\tau}$). The study also referred to the rich products of previous person-fit research, especially those fundamental methodology work between the 1970s and1980s. Because of this, the identification of a person-fit index that can be statistically adapted to the DMM framework rather than building it from the ground up was the primary strategy for the current work.

There are various person-fit and model appropriateness indices from IRT models, but the functioning and usefulness of most indices have not been consistent across different testing conditions. Meijer and Sijtsma (2001) pointed out that test length, trait levels, and misfitting item-score patterns were all influential factors to detection rates of person-fit indices. They also highlighted that the Rasch model framework had developed robust person-fit indices (e.g., *U*, Wright & Stone, 1979; *M*, Molenaar & Hoijtink, 1990), but the Rasch model itself was more restrictive than other IRT models, therefore increasing the starkness of students' misfit. From another perspective, Karabatsos (2003) categorized person-fit statistics into parametric and non-parametric classes. Parametric person-fit refers to the indices that were estimated through model parameters, which are more well-known and easier to understand. In contrast, non-parametric ones were calculated through other testing information. For example, the non-parametric index *C* incorporates a covariance ratio to measure the degree of aberrance of an examinee’s item-response pattern from a perfect pattern (see Equation [4], X_n = examinee *n*’s response vector; *p* = item vectors for proportion correct, X_n^* = examinee *n*’s response vector with correct answers only for the easiest items, Sato, 1975).

$$C = 1 - \frac{Cov(X_{n,} p)}{Cov(X_{n,}^{*} p)} \qquad (4)$$

It has been found that non-parametric person-fit statistics generally perform better than parametric ones (e.g., *U*, *W*, l_0, l_z) in identifying aberrant item-response pattern examinees (Karabatsos, 2003). However, this finding does not imply that researchers should always choose non-parametric approaches over the parametric ones because there are important tradeoffs. Emons et al. (2002) pointed out that the empirical distribution of non-parametric IRT models did not align with the theoretical distribution, and Karabatsos (2003) also warned of the potential tradeoff between power and feasibility of implementation.

In all, none of the model appropriateness statistics can be considered as the best index or generalized to all testing situations, which explains the existence of so many indices in the IRT literature. Thus, it is not necessary to justify why the current study develops a certain type of statistic or index over other available possibilities. Researchers may choose to develop multiple fit indices and use them simultaneously because each statistic provides different information. For example, Sijtsma and Meijer (2001) formulated a non-parametric index, the person response function (PAF). It provides researchers with diagnostic information regarding the types of misfitting in addition to overall model appropriateness information from parametric statistics. Similarly, the DMM framework also allows for multiple TDIs, although one particular formulation was retained after carefully examination in the current study.

Formulating Residual-based DMM Trajectory Deviance Indices

As discussed, DMM has been empirically demonstrated to be an appropriate measurement model for capturing learning capacity. Calculating TDI in DMM can further help validate the capacity estimates of each individual through evaluating the extent to which a students' growth curve deviates from a model-implied trajectory (e.g., Michaelis-Menten curve for TRIAD data, Dumas, McNeish, Sarama & Clements, 2019). The current study initiated the TDI research in DMM with a residual-based method for two main reasons.

First, in addition to three growth parameters (β_{0i}, $\beta_{\mathrm{R}i}$, and $\beta_{\mathrm{A}i}$) in Equation (1), the Michaelis-Menten function also specifies a residual term (d_{ti}), which assesses the difference between Michaelis-Menten model predicted values and the observed values. Individual-specific residuals for each item were successfully used as a component for formulating a parametric fit index in IRT (e.g., *U*, Wright & Stone, 1979; *W*, Wright & Masters, 1982), and therefore (d_{ti}) may be used for the analogous purpose in DMM.

Second, although the empirical example (TRIAD data) used in this study was shown to fit the Michaelis-Menten model best (Dumas, McNeish, Sarama & Clements, 2019), DMM was designed for a diversity of growth shapes in different research settings (Dumas & McNeish, 2017). To reduce restrictions in the usage of the DMM TDI method, the index to be developed should not target a specific growth pattern for DMM. For example, McNeish and his colleagues (McNeish et al., 2019; McNeish & Dumas, 2020) have tested DMM against different growth curves such as the Michaelis-Menten, exponential, Weibull, logistic, and so forth. The exponential function was written as (see Equations of other curves in McNeish et al., 2019):

$$y_{ti} = \beta_{0i} + \left(\beta_{Ci} - \beta_{0i}\right)\exp\left[-\beta_{Ri} Time_{ti}\right] + d_{ti} \qquad (5)$$

As can be seen, these curves all have the within-person residual component for person *i* at time *t* as the Michaelis-Menten function in Equation (1). The definition of the residual parameter (d_{ti}) is the same across all models that have been so far used for DMM research, regardless of their specific functional form. From this perspective, a residual-based index is an ideal choice because the residual-based TDI is not restricted to a certain trajectory, and can therefore be used for a variety of DMM functions.

This dissertation therefore focused on residual-based person-fit indices in the literature. Among those fundamental person-fit methods developed in the late 1970s and 80s, Wright and his colleagues (Wright & Stone, 1979; Wright & Masters, 1982) created a residual-based fit statistic, *Mean Square*, which accumulates squared standardized residuals (Z_{ki}^2) over the answered items (k = 1, . . ., K) for the ith student (i = 1, . . ., N). The Mean Square function of person-fit can be written as:

$$U_i = \sum_{k=1}^{K} Z_{ki}^2 / K \qquad (6)$$

They also invented person-fit (*W*) by weighting the mean square by its variance to reduce the sensitivity to unexpected responses. However, DMM research appeared to prefer such sensitivity of an unweighted index, because a major purpose of developing TDI is to identify students with extreme or surprising scores in their growth curves. It should be also noticed that Both *U* and *W* were considered to be relatively sound among existing IRT person-fit or model appropriateness indices (Meijer & Sijtsma, 2001).

Unweighted DMM Trajectory Deviance Indices. In the DMM context, the residual (d_{ti}) and its variances are both uniquely estimated at each time point, and the scaled ability scores from DMM are analogical to the observed items in IRT models. Thus, these person- and time-specific residuals are suitable components to establish a TDI by adapting the Mean Square function, that is the squared rescaled residual (${D_{ti}}^2$) for the *i*th student (*i* = 1, . . ., *N*) at the *t*th timepoint (*t* = 1, . . ., *T*) are aggregated, and the unweighted DMM trajectory deviance index (TDI$_{\text{U\_Sqr}}$) can be computed as:

$$TDI_{U\_Sqr} = \sum_{t=1}^{T} {D_{ti}}^2/T \tag{7}$$

where

$$D_{ti} = \frac{|d_{ti}|}{\sqrt{Var(|d_{ti}|)}} \tag{8}$$

$Var(d_{ti})$ represents the variance of the residual at the *t*th timepoint. Notably, even the formula (8) for calculating D_{ti} follows the similar way to calculate standardized residuals in the *U* formulation (Wright & Stone, 1979) , D_{ti} is only perceived as re-scaled residuals rather than true standardized values in nonlinear mixed models. This is because the true standardization of DMM residuals using *SAS NLMIXED* requires the incorporation of possible covariance among residuals or covariance among data points (SAS, 2015).

As mentioned, many IRT-based person fit indices have been adapted to other measurement models. However, direct transformations do not always guarantee that those adapted statistics can perform as well as the original ones in the IRT framework. It remains a question whether the unweighted DMM trajectory deviance index has stronger power with aggregated D_{ti} or ${D_{ti}}^2$ overtime. In calculating a person-fit statistic (i.e., response conformity index) for the Cognitive Diagnostic Assessment Model, Cui and Li (2015) found that using absolute values was consistently associated with higher detection rates than using squared values across multiple simulation conditions. Therefore, an unweighted DMM trajectory deviance index (TDI$_{\text{U\_Abs}}$) averaging absolute residual values of D_{ti} across timepoints was also tested:

$$TDI_{U\_\text{Abs}} = \sum_{t=1}^{T} D_{ti}/T \tag{9}$$

Weighted DMM Trajectory Deviance Indices. Both unweighted trajectory deviance indices indicate the averaged trajectory deviance for each person across timepoints, because each time-specific TDI component (D_{ti} or ${D_{ti}}^2$) equally contributes to the TDI$_{\text{U}}$. However, the importance of residuals at different timepoints may vary for the purposes of detecting deviant cases. A critical assumption of DMM models is that a learner-specific trajectory tends to level off as time goes forward (Dumas & McNeish, 2017; Dumas et al., 2020; McNeish & Dumas, 2017), which means asymptote estimates would not be meaningful without meeting this assumption. Based on the assumption, it is reasonable to up-weight the scores at later timepoints and down-weight the early scores in the formula of the weighted trajectory deviance index, and a weighted TDI can be computed as:

$$TDI_{W\_Sqr} = \sum_{t=1}^{T} D_{ti}^{2}/T \tag{10}$$

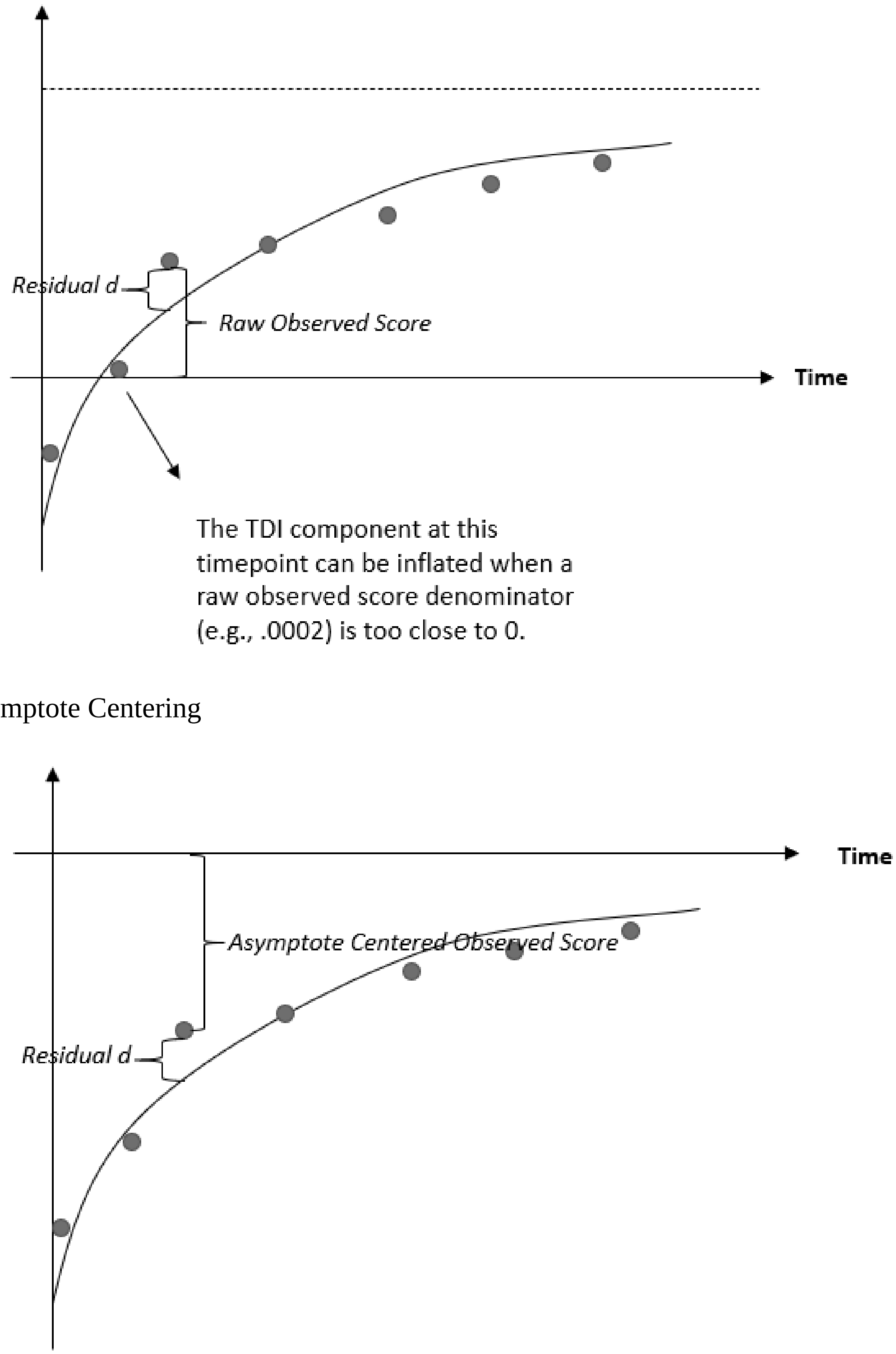
where

$$D_{ti} = \left|\frac{d_{ti}}{Asymptote\ Centered\ Observed\ Score_{ti}}\right| * 100 \tag{11}$$

In comparison to $\mathrm{TDI}_{\mathrm{U}}$, each TDI component ($D_{ti}$) in $\mathrm{TDI}_{\mathrm{W\_Sqr}}$ was computed as the ratio of residual to observed score, which was conceptualized as the percentages of deviance from the observed score for the *i*th student (*i* = 1, . . ., *N*) at the *t*th timepoint (*t* = 1, . . ., *T*). The resulting decimal was multiplied by 100 for scale.

Moreover, the observed scores were asymptote-centered for two reasons. First, TDI components on the original scale can be inflated when an original observed score as the denominator is close to 0 (see Figure 4a), and the occurrence of inflation is hard to predict due to the variability in person-specific growth trajectories. Second, the asymptote-centered observed scores gradually decrease as time goes (see Figure 4b), which results in larger deviance components at later timepoints. In other words, later deviance components are up-weighted. Other centering approaches that weight deviance components differently have also been considered. For example, centering observed scores around the intercept for each person can up-weight early behavior and down-weight late behavior (see Figure 4c). However, such a weighting strategy cannot address the purpose of a weighted TDI, that is, increasing TDI sensitivity to detect cases whose growth trajectories did not level off.

(a) Original Scale



The TDI component at this timepoint can be inflated when a raw observed score denominator (e.g., .0002) is too close to 0.

(b) Asymptote Centering

(c) Intercept Centering

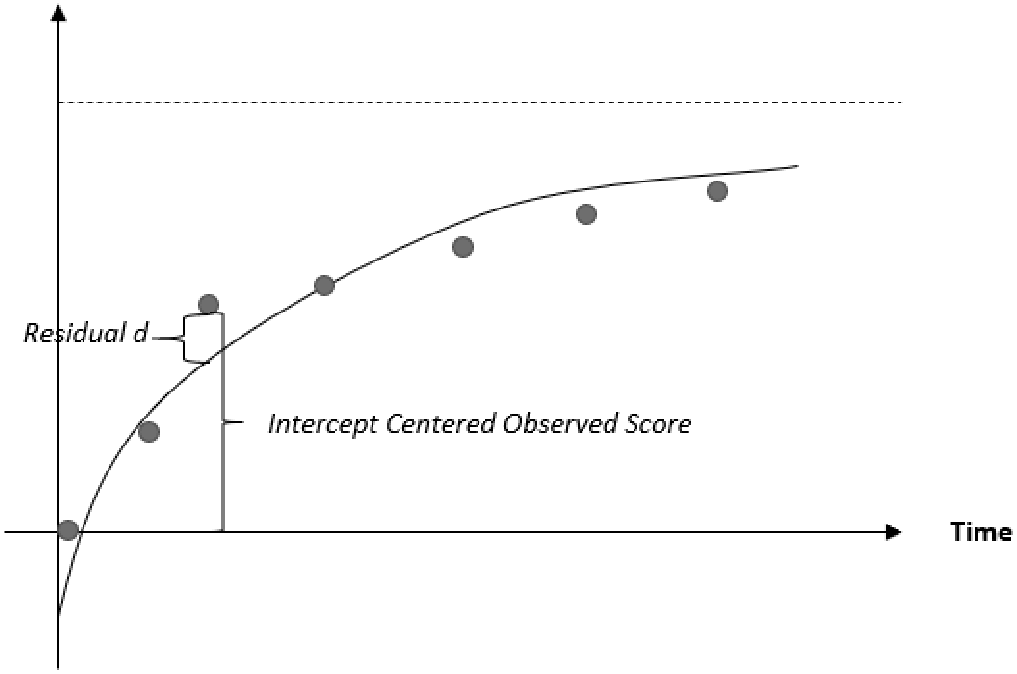


Figure 4. Weighted TDI Components on Different Scales

Like formulating the unweighted trajectory deviance indices, the study also examined the TDI$_{W\_Abs}$ in which the absolute values of deviance components rather than squared values are aggregated:

$$TDI_{W\_Abs} = \sum_{t=1}^{T} D_{ti}/T \qquad (12)$$

The weighting strategy showing in formula [11] was also applied to TDI$_{W\_Abs}$. As a result, a total of four indices have been formulated, but only the one showing appropriate distributional properties and efficacy in detection will be retained at the end of the study.

All these TDIs are designed to tolerate missing values in a similar way as many IRT statistics: when there is a missing score(s) at the *t*th timepoint, this score just does not contribute to the mean square function. This feature could be particularly useful to

DMM because missing data is a common issue in the analysis of longitudinal data. However, a high ratio of missing to valid scores could still significantly decrease the accuracy of estimating person fit and other basic parameters.

This chapter demonstrates that conventional static measurement practices have contributed to the sustained achievement gaps from multiple aspects, and they are not appropriate to assess student learning potential. In contrast, studies have shown that DMM can generate asymptotical capacity estimates and improve the consequential validity of educational measurement, but it is still a continuing task to refine this newly invented measurement framework. Four TDIs aiming to detect aberrant cases to DMM models has been formulated, and the next chapter focuses on the methodological details in examining and comparing the performances of different TDIs.

Chapter Three: Method

Data Source

The empirical example used in this project was the TRIAD cluster-randomized experiment data (Sarama & Clements, 2013; Sarama et al., 2008), which contained seven waves of mathematics test scores: pre and post of PreK, Kindergarten, 1st grade, 3rd grade, 4th grade, and 5th grade. All test scores were generated through IRT models, and thus all scores align on a single continuous vertical scale. The TRIAD data were used for addressing research goals in both stages of the project.

Participants

The TRIAD sample originally consisted of 1,375 students from 42 schools and 106 classrooms in two cities (Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Clements et al., 2013). Schools were randomly assigned to one of the three conditions: *Building Blocks* (BB) intervention only at the PreK year, BB intervention with follow-through in the kindergarten and first-grade years (BB with FT or BBFT), and business-as-usual (BAU; control). At the post-test in the PreK year, there was about 5% (n = 70) attrition of participants, but the attrition was not related to their entering performance.

The final analytic sample was 1,305 students aged from 44 to 64 months (M = 52.06, SD = 4.09) at the pre-test in PreK. Of these participants, 51% (n = 664) were female, 24.5% (*n* = 319) reported themselves as bilingual, 10% (*n* =130) had documented special needs, and 66.7% (*n* = 871) received Free/Reduced lunch. Regarding race and ethnicity, over half of children (53.3%, *n* = 695) were African American, 21.6% (*n* = 282) were Hispanic, 18.9% (*n* = 247) were White, 3.7% (*n* =48) were Asian American, 1.8% (*n* =24) were Native American, and .6% (*n* = 8) reported their ethnicity as Other. The participant attrition rate was 13.74% (*n* = 179) at the end of the first-grade year, and it was shown that the attrition was unrelated to the demographics (Clements et al., 2020). In the analytic sample, 605 (46.36%) students have complete assessment data across all seven timepoints. The follow-up data after the 1st grade year were collected for evaluating the persistence of the BB effects, and there was no additional intervention implemented.

Assessment

Core mathematical knowledge of students was assessed with the Research-based Early Math Assessment (REMA, Clements et al., 2008). The measure was designed for students from ages 3- to 8-years-old with standardized administration, videotaping, scoring, and coding procedures. The REMA measurement tool reflects developmental progressions of early mathematics and covers a wide range of topics: verbal counting, object counting, subitizing, number comparison, number sequencing, connection of numerals to quantities, number composition, decomposition, adding and subtracting, place value, shape recognition, shape composition and decomposition, congruence, construction of shapes, spatial imagery, geometric measurement, patterning, and reasoning (Clements et al., 2013). The data were collected using the Full version of

REMA with 225 items, and validity evidence was demonstrated in multiple aspects (see content, construct, and concurrent validity in Clements et al., 2008).

All REMA items are ordered by item difficulty, and each participant stops the test after giving incorrect answers to four consecutive items. All assessors received formal training on administration, videotaping, scoring, and coding procedures, and each was required to be certificated (98-100% error-free delivery) before assessing any participants. In addition to item correctness, the REMA also tests the strategy use of students. Assessors observed, videotaped, recorded, and noted students' strategies to solve math problems. Based on the taped videos as well as field notes from assessors, a trained coder coded each strategy item. Continuous feedback was delivered from an expert coder to trained coders for calibration purposes (one tape per coder per week). In the scoring stage, all strategy codes were recoded based on the sophistication levels, and both strategy levels and item correctness were included in the scoring model. Latent mathematics scores were generated through IRT models, and the reliability of total scores was high: .93 to .94 in an earlier study (Clements et al., 2008) and .92 for the TRIAD data (Clements et al., 2011).

Intervention

One important characteristic on which dynamic testing differs from static testing is the integrated instruction for supporting student growth between measurement occasions (Dumas et al., 2020). The TRIAD project implemented the *Building Blocks* curriculum, which was structured on learning trajectories and featured with the use of a comprehensive Curriculum Research Framework (Clements et al., 2013). Research showed that the short-term BB significantly improved students' mathematics competency (Clements et al., 2013; Dumas, McNeish, Sarama & Clements, 2019), and the detected effect sizes (g = .72) were above medium level. Moreover, the BB treatment effect diminished over time (Bailey et al., 2016), and such a pattern has been frequently observed for early childhood interventions (Jenkins et al., 2018; Kang et al., 2019). As a result, students who received the intervention or not may demonstrate very different growth paths.

Data Analysis

This project consists of two studies (simulation and real-data application), and the specific analysis procedures for each part are delineated below.

Simulation Study

The simulation analysis was performed in two steps for different purposes. The first step involved resampling with replacement to identify distributional characteristics and find empirical critical values (e.g., α = .05) of TDI statistics, and the second step investigated and compared the effectiveness of the formulated TDIs following a 3 × 3 simulation design.

Distributional Properties and Empirical Critical Values. In the first step, bootstrapping, a resampling approach with replacement[2], was utilized to obtain the distributional characteristics of the formulated trajectory deviance statistics ($\text{TDI}_{\text{U\_Abs}}$, $\text{TDI}_{\text{U\_Sqr}}$, $\text{TDI}_{\text{W\_Abs}}$, and $\text{TDI}_{\text{W\_Sqr}}$). Bootstrapping is a special type of Monte Carlo

[2] Replacement allows the same case to be sampled multiple times, which can induce inter-replication variability.

simulation based on given data without specifying any underlying data generating process. To avoid “garbage in and garbage out” and ensure the representativeness of the simulated data to a real data context, the well-researched TRIAD data (Bailey et al., 2016; Clements et al., 2013; Clements et al., 2020; Dumas, McNeish, Sarama & Clements, 2019; Sarama et al., 2008) produced population estimates for the data generation model. The TRIAD data has shown fit to the Michaelis-Menten model in an earlier study (Dumas, McNeish, Sarama & Clements, 2019), so the current study directly applied the Michaelis-Menten function to model seven waves of REMA scores. The model used a homogeneous diagonal residual error structure and an unstructured random effect covariance structure to achieve parameter estimates. In order to reflect the timing of TRIAD data collection, the time scores (*t*) in the slope factor as set to 0, 0.5, 1, 2, 4, 5 and 6 for pre and post of PreK, Kindergarten, 1st grade, 3rd grade, 4th grade and 5th grade measures, respectively.

For each TDI statistic, the simulation was conducted within two sample size conditions: 605 and 300. The 605 was the actual number of children with complete assessment data across seven timepoints. The 300 sample size condition was included to investigate the properties of TDIs when the sample size was smaller and the model was more difficult to estimate, which follows the methodological design of developing DMM reliability methods (McNeish & Dumas, 2018). Each simulation condition in this step was replicated 100 times. The simulation analyses were conducted in *SAS 9.4* using *PROC NLMIXED* (SAS, 2015), and models were estimated with Gaussian quadrature with five quadrature points and double dogleg optimization (max iteration = 5000). To save computational times, this study used virtual machines to simultaneously run multiple SAS programs (smaller number of replications within each) via three computers, and the results generated from each device were merged later. The total time to finish the simulation was about five days.

Evaluating the Effectiveness of TDIs. Empirical critical values found in the previous simulation were then used for detecting deviant (i.e., misfitting) growth trajectories in the second part of the simulation analysis. The effectiveness of four TDIs were evaluated according to three statistical indices: detection rate, type I error rate, and Rand index (Rand, 1971). Specifically, the detection rate was calculated as the averaged proportion of aberrant growth trajectories correctly identified as aberrant across all simulation replications, and type I error rate was calculated as the averaged proportion of fitting growth trajectories identified as aberrant across all replications. Rand index originally was designed to assess similarities between data clusters (Rand, 1971), and it has been extended to use as a measure of the proportion of correct decisions via the following algorithm:

$$RAND = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

The second simulation randomly generated growth curves via different functions instead of using the resampling approach described in the previous section. However, the data generation model was still based on parameter estimates of TRIAD data to ensure that values are representative of real-world data. For each formulated TDI, nine conditions were simulated following a 3 × 3 design. The first factor featured three types of aberrant growth trajectories to the Michaelis-Menten trajectory (inverted J-shape):

linear growth, logistic growth (S-shape), and accelerative or exponential growth (J-shape). Each type of aberrant growth curve is commonly observed in educational settings, and a TDI needs to be capable of identifying such deviant patterns in further research. However, the power in detecting different aberrant growth may vary due to the inherent differences among those curves. For example, the exponential growth curves (J-shape) should be the easiest ones to detect because they are the most different from Michaelis-Menten trajectories (inverted J-shape).

The study also aimed to understand the robustness of the formulated TDIs with different percentages of deviant cases in the data for modeling, and therefore three percentages of cases showing aberrant growth patterns (5%, 10%, or 15%) in the generated datasets were simulated. The percentage of deviant cases is a key factor that can affect the detection ability of many person-fit statistics. Karabatsos (2003) compared 36 person-fit statistics and found that detection rates generally decreased as the percentage of aberrant respondents increased. The current project did not simulate higher percentage conditions (e.g., 20%, 30%) because applying modeling strategies such as multi-group modeling techniques is more appropriate to accommodate very high portions of aberrant cases.

Fifty replications were conducted for each condition, which generated 450 datasets in total at this step. Each dataset has 605 simulated cases, which is the number of TRIAD cases with completed data. Four TDIs were calculated within all simulated datasets and averaged values of three efficacy indices (i.e., Detection rate, Type I error, Rand) were computed over all of the replications.

Real-Data Application

TDI measures the magnitude of deviance from student-specific to the overall growth curve (e.g., inverted-J curve). Based on the results of simulation analyses, the most effective and robust TDI was retained. To demonstrate how this index can be incorporated into DMM framework, the second part of the study applied it to explore the contributing factors to the deviance of growth in early mathematics. The hierarchical linear modeling (HLM) approach was utilized to investigate the effects of both classroom-level and student-level predictors (Raudenbush & Bryk, 2002).

Classroom-level Predictors. The TRIAD intervention group received the BB curriculum. It has been found that students from the intervention group performed significantly better than those from the control groups in the immediate post-test (Clements et al., 2013; Clements et al., 2020). Meanwhile, the intervention group was shown to have significantly higher learning rates from a long-term perspective (Dumas, McNeish, Sarama & Clements, 2019). Both findings indicate that the BB intervention may change student growth from a natural trajectory, but it remains a question whether the intervention would significantly impact the magnitude of trajectory deviance (i.e., the accuracy of estimates).

Student-level Predictors. In addition to the intervention effect, this study also examined and controlled several student demographic variables, including special education status, bilingual status, SES, gender, and age. Those variables have been often associated with mathematical growth and learning (e.g., Clements et al., 2020; Duncan & Magnuson, 2012; Kenney-Benson et al., 2006; Siegler & Booth, 2004). In the current analysis, SES was indicated by whether a student was eligible to receive Free/reduced or full-paid lunch, and age was grand mean centered for interpretation purposes.

TDI was treated as a continuous outcome in the HLM analysis above, but TDI or other developed model appropriateness statistics have been often used dichotomously in research practices (e.g., evaluate whether participants fitted the model, Dong et al., 2020; Meijer & Sijtsma, 2001). Similarly, TDI may be applied to evaluate whether each growth trajectory is substantially deviant from the model implied curve according to the recommended critical values. For this purpose, the established TDI was applied to identify aberrant growth curves in the TRIAD data. Meanwhile, I also compared the conditional reliability of DMM with and without the deviant cases.

Chapter Four: Results

This chapter presents two simulation analyses and the application of TDI into the TRIAD early mathematics data. In correspondence to the analytic procedures described above, results were organized into three sections focusing on the distributional properties of the formulated TDIs, the effectiveness of the TDIs, and an empirical application of the TDI, respectively.

Distributional Properties and Empirical Critical Values of TDIs

A simulation study using bootstrapping techniques was first conducted to examine the distributional properties (i.e., mean, standard deviation, skewness, kurtosis[3]) of four TDI candidates. Empirical critical values later used for detecting aberrant trajectories in the following analyses were found at the significance level of .05. The overall convergence rates of bootstrapped datasets were high: 99 of 100 models in the 605 sample size condition and 98 of 100 models in the 300 sample size condition successfully converged.

Descriptive Statistics

Table 1 presents the mean and standard deviation for the distributional properties and critical values of each TDI across two sample size conditions. In the 605 sample size

[3] All kurtosis values reported in this study were excess kurtosis.

Table 1.

Distributional Properties and Empirical Critical Values of TDIs

	Mean		**SD**		**Skewness**		**Kurtosis**		**Critical Value**	
	M	SD	M	SD	M	SD	M	SD	M	SD
Sample size = 605										
$\text{TDI}_{\text{U\_Abs}}$	1.26	.01	.42	.01	.57	.10	.30	.26	2.08	.04
$\text{TDI}_{\text{U\_Sqr}}$	2.60	.04	1.80	.07	2.00	.43	7.77	4.25	6.09	.23
$\text{TDI}_{\text{W\_Abs}}$	16.68	5.27	57.05	90.40	12.18	5.23	197.56	153.85	30.28	9.12
$\text{TDI}_{\text{W\_Sqr}}$	79637.23	328038.66	1643896.57	7700290.23	17.14	5.05	330.65	178.17	3209.18	5370.47
Sample size = 300										
$\text{TDI}_{\text{U\_Abs}}$	1.27	.03	.42	.02	.54	.14	.21	.34	2.08	.05
$\text{TDI}_{\text{U\_Sqr}}$	2.61	.07	1.78	.11	1.78	.57	5.74	4.92	6.09	.29
$\text{TDI}_{\text{W\_Abs}}$	18.19	13.61	61.93	137.37	8.97	3.78	106.58	80.25	31.61	8.65
$\text{TDI}_{\text{W\_Sqr}}$	158562.48	1057663.24	1793169.64	10648541.64	12.63	3.43	177.50	87.86	3610.03	3451.92

Notes. Each cell summarized the results of 100 replications.

condition, the averaged skewness of $\mathrm{TDI}_{\mathrm{U\_Abs}}$ across 100 replications was .57 (*SD* = .10), and the averaged kurtosis was .30 (*SD* = .26). Both values were within the range of -1 to 1, which means $\mathrm{TDI}_{\mathrm{U\_Abs}}$ followed a normal distribution. The averaged skewness of .54 (*SD* = .14) and kurtosis of .21 (*SD* = .34) in the 300 sample size condition indicated the same conclusion. In contrast, the distribution of $\mathrm{TDI}_{\mathrm{U\_Sqr}}$ was positively skewed (605 sample size: *M* = 2.00, *SD* = .43; 300 sample size: *M* = 1.78, *SD* = .57) and leptokurtic (605 sample size: *M* = 7.77, *SD* = 4.25; 300 sample size: *M* = 5.74, *SD* = 4.92) under both conditions.

As can be seen in Table 1, the two weighted TDI statistics were severely skewed and leptokurtic. The very large mean kurtosis across conditions (ranging from 106.58 to 330.65) implies frequent occurrences of outliers in $\mathrm{TDI}_{\mathrm{W\_Abs}}$ and TDI $_{\mathrm{W\_Sqr}}$. Furthermore, the corresponding standard deviations were also very large (ranging from 80.25 to 178.17), which indicates the distributions of the weighted TDIs varied much across the replications within each sample size condition.

Pooling the Critical Values across Sample Sizes

To find the critical value for each TDI at the significance level of .05, the averaged fifth percentile values were calculated across 100 replications in each condition. The critical values of both unweighted TDIs ($\mathrm{TDI}_{\mathrm{U\_Abs}}$ critical value = 2.08; $\mathrm{TDI}_{\mathrm{U\_Sqr}}$ critical value = 6.09) were identical across the two sample size conditions. The averaged critical values of $\mathrm{TDI}_{\mathrm{W\_Abs}}$ (*M* = 30. 28, *SD* = 9.12) and $\mathrm{TDI}_{\mathrm{W\_Sqr}}$ (*M* = 3209.18, *SD* = 5370.47) were smaller in the 605 condition than the 300 condition ($\mathrm{TDI}_{\mathrm{W\_Abs}}$ critical value: *M* = 31.61, *SD* = 8.65; $\mathrm{TDI}_{\mathrm{W\_Sqr}}$ critical value: *M* = 3610.03, *SD* = 3451.92), but the differences were not statistically significant. Therefore, the results were pooled across the 200 replications to determine the final critical values for the later examination. Considering the convenience of future application of TDI, I only kept one decimal in those pooled means in order to facilitate later comparison to the cut-off values: $\text{TDI}_{\text{U\_Abs}}$ critical value = 2.1; $\text{TDI}_{\text{U\_Sqr}}$ critical value = 6.1; $\text{TDI}_{\text{W\_Abs}}$ critical value = 31.0; $\text{TDI}_{\text{W\_Sqr}}$ critical value = 3397.0. These cut-off values were on their own scales to reflect their unique meanings based on their formulations. For example, a $\text{TDI}_{\text{W\_Abs}}$ of 31.0 indicates 31% averaged deviance from the growth trajectory across timepoints. Nevertheless, the outcomes (i.e., power, Type-I error rates, Rand values) in the evaluation of four TDIs will be on the same scale and directly comparable.

Evaluating Effectiveness of Formulated TDIs

A second simulation with a 3 (types of aberrant trajectories) × 3 (portions of aberrant trajectories) design was performed to evaluate the effectiveness of that four candidate TDIs. The sample size was not included as a factor here because of its negligible influences on the TDIs shown in the previous simulation. Each generated dataset contained a total of 605 cases: 30 aberrant and 575 fitting trajectories for the 5 % aberrant condition; 60 aberrant and 545 fitting trajectories for the 10 % aberrant condition; 90 aberrant and 515 fitting trajectories for the 15 % aberrant condition.

To ensure representativeness of real data in the generated datasets, parameters in the functions (i.e., linear, J-shaped, S-shaped, inverted-J shaped) were all based on the TRIAD early mathematics data. For example, describing the growth between scores at the initial and last timepoints of TRIAD cases using linear models resulted in an averaged slope of .79 (*SD* = .17). Then, a random slope (S_i) from a normal distribution (*M* = .79, *SD* =.17) was assigned to a linear function ($Math_{ti} = S_i \times t + \beta_{0i}$) to generate linear growth data for the simulation datasets. β_{0i} was also a random initial value from the TRIAD-based normal distribution. Similarly, modeling the growth between the initial and last scores of TRIAD cases using a J-shaped function ($Math_{ti} = \beta_{0i} + X_i^t$ -1) resulted in an averaged base number (*X*) of 1.34 (*SD* =.04). Two random parameters derived from corresponding distributions were then assigned to each case to generate the J-shaped growth data. To produce the S-shaped growth data, four random parameters were required in terms of its function: $Math_{ti} = \beta_{Li} + \frac{\beta_{Ai} - \beta_{Li}}{[1+e^{-\beta_{MSi}(t-\beta_{Ri})}]}$. In addition to the common capacity (β_A) and learning rate (β_R) parameters in most DMM models, the S-shaped function also comes with two unique logistic growth parameters (β_L, lower asymptote; β_{MS}, the slope at midpoint). The distributions of these four parameters were all based on the estimates from TRIAD cases. Besides three types of aberrant trajectories, growth data in the shape of the model implied curve (inverted J) was produced via the Michaelis-Menten model with random growth parameters.

General Patterns of the TDIs Functioning across Simulation Conditions

A total of 450 datasets (3 × 3 × 50) following the simulation design were generated and imported into the Michaelis-Menten model. The convergence rate was still high: only 17 (3.8%) of 450 DMM models did not converge. The frequencies of non-convergence had a weak relation to the types of aberrant trajectories (point biserial *r* = .09) but a moderate relation (point biserial *r* = .38) to the percentages of deviant cases within the dataset. The time- and person-specific residuals from each model were used in

Table 2.

Comparisons of TDIs across Simulation Conditions

	Linear growth						Exponential Growth (J-shape)						Logistic Growth (S-shape)					
	Detection rate		Type I error		Rand		Detection rate		Type I error		Rand		Detection rate		Type I error		Rand	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
5 % aberrant growth trajectories																		
$\mathrm{TDI}_{\mathrm{U\_Abs}}$	.82	.22	.04	.03	.95	.03	.96	.07	.02	.01	.98	.01	.40	.31	.02	.03	.95	.03
$\mathrm{TDI}_{\mathrm{U\_Sqr}}$	.77	.23	.03	.02	.96	.02	.92	.09	.02	.01	.98	.01	.40	.32	.02	.02	.95	.02
$\mathrm{TDI}_{\mathrm{W\_Abs}}$	.18	.15	.00	.00	.96	.01	.87	.16	.00	.00	.99	.01	.03	.04	.01	.01	.95	.01
$\mathrm{TDI}_{\mathrm{W\_Sqr}}$	.14	.12	.00	.00	.96	.01	.85	.16	.00	.00	.99	.01	.02	.03	.01	.01	.95	.01
10 % aberrant growth trajectories																		
$\mathrm{TDI}_{\mathrm{U\_Abs}}$	.77	.10	.04	.02	.94	.01	.85	.11	.02	.01	.97	.01	.31	.26	.02	.02	.92	.03
$\mathrm{TDI}_{\mathrm{U\_Sqr}}$	.65	.11	.03	.01	.94	.01	.75	.11	.02	.01	.96	.01	.30	.24	.01	.01	.92	.03
$\mathrm{TDI}_{\mathrm{W\_Abs}}$	.05	.05	.00	.00	.91	.00	.77	.12	.00	.00	.98	.01	.02	.03	.01	.01	.90	.00
$\mathrm{TDI}_{\mathrm{W\_Sqr}}$	.04	.04	.00	.00	.90	.00	.75	.13	.00	.00	.97	.01	.01	.03	.00	.01	.90	.00
15 % aberrant growth trajectories																		
$\mathrm{TDI}_{\mathrm{U\_Abs}}$	.63	.15	.02	.01	.93	.03	.66	.13	.01	.01	.94	.02	.15	.16	.01	.01	.86	.03
$\mathrm{TDI}_{\mathrm{U\_Sqr}}$	.47	.14	.02	.01	.91	.02	.53	.12	.01	.01	.92	.02	.15	.14	.01	.01	.86	.03
$\mathrm{TDI}_{\mathrm{W\_Abs}}$	.05	.06	.00	.00	.86	.01	.69	.11	.00	.00	.95	.02	.04	.05	.01	.02	.85	.01
$\mathrm{TDI}_{\mathrm{W\_Sqr}}$	.04	.05	.00	.00	.86	.01	.68	.13	.00	.00	.95	.02	.03	.04	.01	.02	.85	.01

Notes. Each cell summarized the results of 50 replications, and both extreme and non-extreme values were included.

the calculation of four TDIs, and calculated critical values in the first simulation were then used to evaluate the effectiveness of each TDI. Table 2 presents means and standard deviations of detection power, Type I error, and Rand index for each TDI across simulation conditions.

Detection Power. The detection power of the TDIs was calculated as the percentages of correctly identifying aberrant trajectories. In general, TDIs showed the highest power in detecting J-shape curves (*M* ranging from .53 to .96), moderate power in detecting linear growth (*M* ranging from.04 to .82), and lowest power in detecting S-shape curve (*M* ranging from.02 to .40). This pattern was consistent for all four TDIs, which implies that the difficulties of detecting aberrant trajectories varied on the forms of growth trajectories. Another common pattern from Table 2 was that the detection power decreased as the percentages of aberrant growth increased in the simulation data. This finding was consistent with a previous investigation about the detection power of 36 person-fit statistics (Karabatsos, 2003).

Type I Error. The detection power only reveals one aspect of the efficacy of four TDIs, which is not sufficient to completely describe the functioning of the statistics. This is because an index with a very liberal cutoff may lead to high detection power but meanwhile inflate Type-I error (i.e., false positives). Therefore, the Type I error rates of four TDI candidates were also calculated and displayed in Table 2. The Type I error rates of all TDIs were consistently small across all replications within different simulation conditions: *M* ranged from 0 to 4%, and *SD* ranged from 0 to .03, which implies that all four TDIs were not oversensitive in their detection. Notably, though smaller type I error rates are preferred in general, the suggested range of this value is α ± 1/2α (i.e., 025 to .075, Bradley, 1978; Hancock, Lawrence & Nevitt, 2000).

Rand Index. Despite that detection power and Type I error are useful criteria to understand a fit statistic, it can still be hard to make decisions regarding which TDI outperformed others because a less powerful index may demonstrate lower Type I error rates, and vice versa. Thus, the Rand index was used as a composite criteria in this evaluation to determine which TDI index had the optimal properties. The calculated Rand values ranged from .85 to .98, which means most decisions made based by TDIs and their critical values were correct for all four candidate indices. Similar to the power, it was also observed that Rand values decreased as the percentages of cases displaying aberrant growth increased in the tested datasets.

Finding the Best Functioning TDI among the Four Candidates

Differences in the effectiveness of the four TDIs were further examined to find the best functioning TDI. According to the results above, all four indices showed small Type I error rates and satisfying Rand values across nine simulation conditions, which suggests that the detection power should be the critical determining criteria in this case. In Table 2, the power of both weighted TDIs was weak (*M* ranged from .01 to .14) in all conditions except the condition with 15% exponential growth curves. Therefore, it appeared that TDI_{W_Abs} and TDI_{W_Sqr} were not effective in detecting aberrant trajectories in most cases.

As for the two unweighted TDIs, TDI_{U_Abs} showed equal or higher power than TDI_{U_Sqr} across all the conditions, so TDI_{U_Abs} appeared to be better than TDI_{U_Sqr}. The averaged power of TDI$_{U\_Abs}$ in identifying aberrant cases with linear growth, exponential growth, and logistic growth was .74 (*SD* = .18), .82 (*SD* = .16), and .29 (*SD* = .31), respectively. Although the TDI$_{U\_Abs}$ was formulated as the averaged deviance across timepoints, it performed particularly well in detecting types of growth with no leveling-off process (i.e., a critical assumption of DMM).

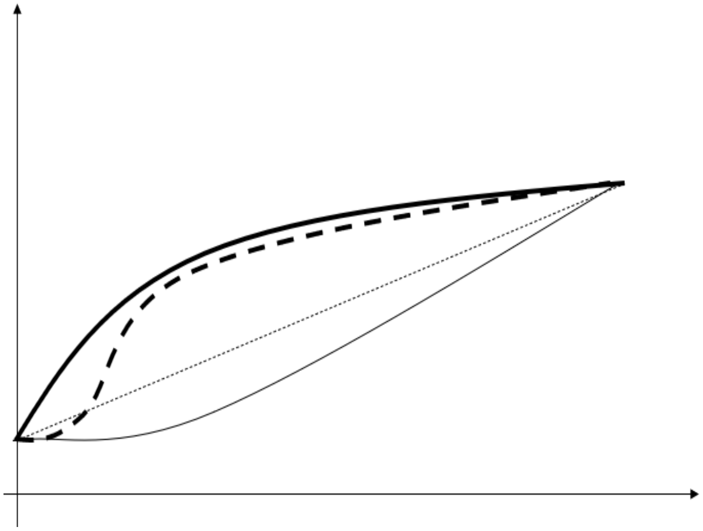


Figure 5. Different Forms of Hypothetical Growth Trajectories

Notably, the power to detect logistic growth curves was relatively low, but this would not invalidate TDI$_{U\_Abs}$ in effect. As displayed in Figure 5, a fitting curve (solid curve in bold) is much more similar to a logistic growth curve (dashed in bold) than the other growth trajectories. More importantly, the S-shaped logistic growth also leveled off as time passes, which meets the essential assumption of using DMM to generate asymptotic capacity estimates (Dumas & McNeish, 2017; Dumas et al., 2020; McNeish & Dumas, 2017). In recent work, McNeish et al. (2019) also empirically showed that the Michaelis-Menten model and the logistic model both fitted the same data and showed very close mean square errors. Taken together, using the Michaelis-Menten model to describe part of the logistic growth data can be supported both theoretically and empirically, and the relatively low power of $\text{TDI}_{\text{U\_Abs}}$ in detecting the logistic curves was therefore acceptable within the DMM framework.

In addition to the highest power, $\text{TDI}_{\text{U\_Abs}}$ was the only index with a stable normal distribution among the investigated TDIs, which makes it an ideal continuous measure for the magnitude of deviance from student-specific to the model-implied growth curve. For all these reasons, $\text{TDI}_{\text{U\_Abs}}$ was determined to be the final TDI formulation. Notably, the empirical critical value (i.e., 2.1) of $\text{TDI}_{\text{U\_Abs}}$ should not be perceived as a universal rule of thumb or hard cut-off in future applications. Follow-up studies may consider establishing a generalizable cut-off and its associated confidence interval based on additional simulation work.

Empirical Example: Applying the TDI into the TRIAD data

In the empirical example, the developed TDI was then applied to explore the influences of the BB intervention and student demographics (i.e., special education status, bilingual status, SES, gender, and age) on the deviance from a typical growth trajectory in early mathematics. Meanwhile, we also examined whether the validity of DMM score use and interpretation can be improved after incorporating the TDI.

Contributing Factors to the Deviance of Growth in Early Mathematics

The TRIAD data was modeled in the Michaelis-Menten function, and each student received a TDI value calculated by the formula [9]. Figure 6 displays the histogram of TDI Values in the TRIAD Data, and it shows a clear normal distribution (*M* = 1.26, *SD* = .42, *skewness* = .59, and *kurtosis* = .33). The following HLM analysis then treated TDI as a continuous outcome variable. A null HLM model with no predictors was tested first. An ICC of .02 (*SE* = .03) indicated that the clustering effect on TDI was weak (less than .05, Raudenbush & Bryk, 2002), so it was not necessary to conduct a multi-level model in the prediction of TDI. However, this empirical example still followed the original analysis plan of running HLM analysis, in order to explicitly model classroom and student-level predictors.

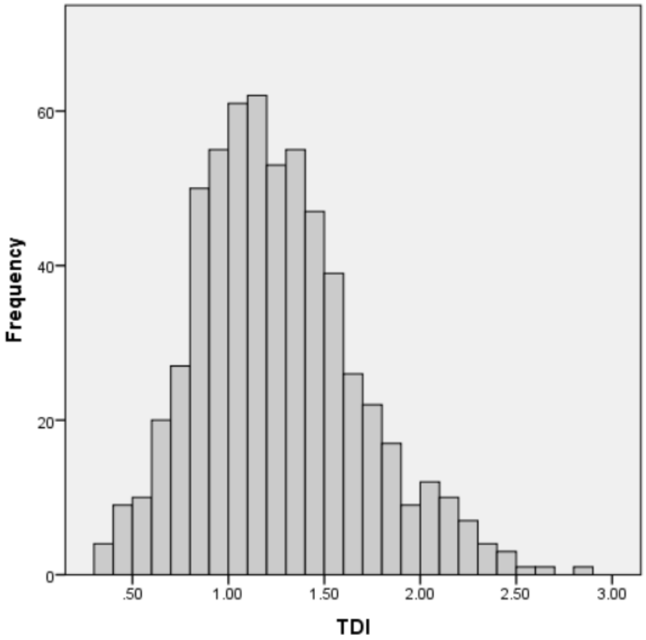


Figure 6. Histogram of TDI Values in the TRIAD Data

Table 3 summarized the HLM model predicting TDI values with both classroom- and student-level predictors. After controlling the student-level predictors, the intervention status was not associated with the trajectory deviance (*B* = -.08, *SE* = .04, *z* = -1.96, *p* = .05), which means there was no bias in the accuracy of asymptote estimates between intervention groups. This finding could further support the asymptote-related conclusions in a previous DMM study (Dumas et al., 2019) using the same data.

Table 3.
Multi-level Linear Regression Predicting TDI

Predictors	**Coefficient**	**SE**	**z**	**p**
Classroom-level Predictor				
Intervention	-.08	.04	-1.96	.05
Student-level Predictors				
Age	.00	.00	.18	.86
Special Ed	.12	.06	1.90	.06
Gender	-.03	.04	-.79	.43
Bilingual	.09	.04	2.13	$.03^{*}$
SES	.05	.05	1.06	.29

Notes. * $p < .05$.

Regarding the student-level predictors, bilingual students had significantly higher TDI values ($B = .09$, $SE = .04$, $z = 2.13$, $p = .03$), which means they deviated more from model-implied growth curves than non-bilingual students. In the next section, multiple sources of evidence were checked to verify whether these deviances would substantively invalidate the use of DMM capacity estimates.

Post-hoc Validity Check using Bilingual Status as an Example

Validity is a property of the proposed interpretation and use of scores (Kane, 2013). Asymptote scores within the DMM framework have been interpreted as the quantity of learning capacity in a certain domain. The average growth trajectories of bilingual and non-bilingual students were first plotted to examine the form of their shapes and whether their asymptotes can be meaningfully interpreted. From Figure 7, both curves were in a general J-shape and leveled-off in the later timepoints, which supports the validity of interpreting DMM asymptotes as learning capacities for both groups of students.

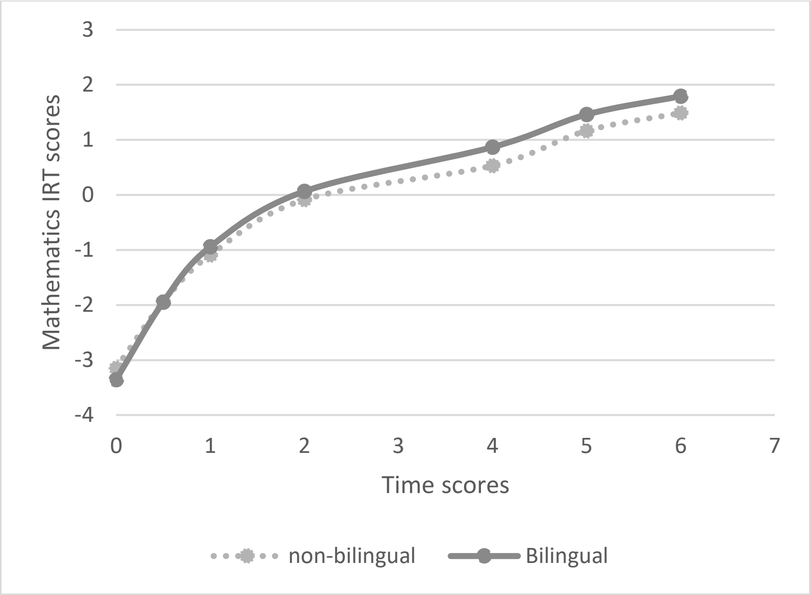


Figure 7. Growth Trajectories of Bilingual and Non-bilingual Students

Then, a total of 27 (4%) students were flagged based on the empirical critical value (i.e., TDI over 2.1 indicates an aberrant trajectory). It was found that the bilingual characteristic was not associated with being identified as aberrant (χ^2 = .71, *p* =.40). This finding implies that bilingual students did not as a group significantly disproportionately misfit to the model-implied trajectory, though they did have higher TDIs than their counterparts on average. Finally, the relation between asymptotic capacity and TDI was checked via a scatterplot (see Figure 8) and Pearson correlation (*r* =.04, *p* =.31). Both results indicated no relation between the two DMM quantities.

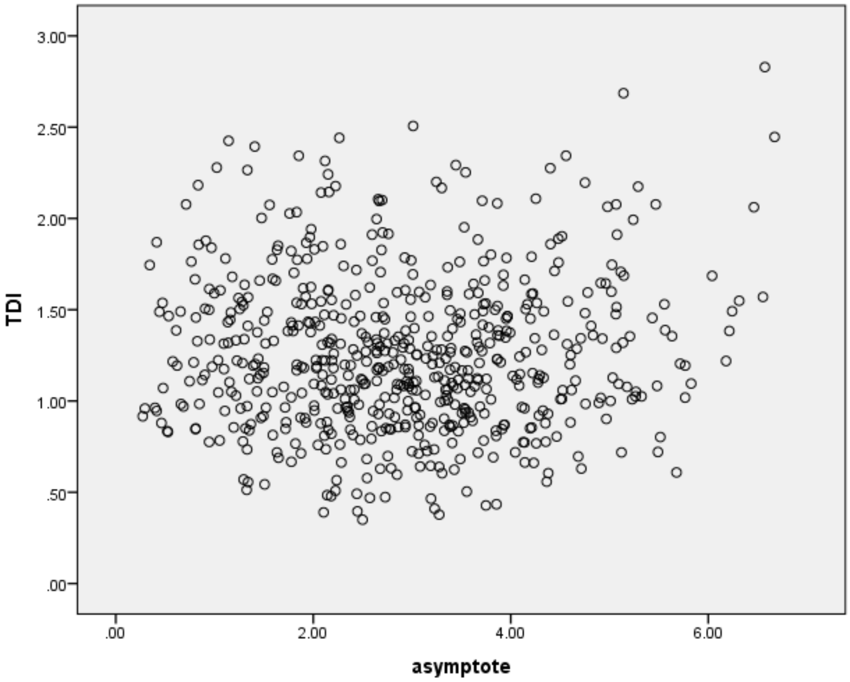


Figure 8. Scatterplot of TDI and Asymptote

In sum, the identified influences of bilingual status on TDI neither invalidated the use of asymptote estimates nor skewed related findings. With these kinds of validity check procedures, researchers would be more confident about the findings derived from DMM analysis. In this empirical example, TRIAD bilingual students were found to enter with significantly lower performance (t = -2.71, p = .007), catch up at the end of pre-k (t =.02, p =.984), gradually expand their learning in later years, and meanwhile show significantly higher learning capacities than their counterparts in mathematics learning (t = 3.27, p = .001). This dissertation did not delineate these findings because it was beyond the scope of the current study.

In this chapter, the distributional properties and effectiveness of four formulated TDIs were examined and compared. $\text{TDI}_{\text{U\_Abs}}$ has a normal distribution and shown the best functions in detecting various aberrant growth trajectories. Thus, the $\text{TDI}_{\text{U\_Abs}}$ was determined to be the formal trajectory deviance index for DMM. Moreover, I illustrated some uses of TDI by applying the index to the TRIAD data. The next chapter discusses major findings in detail and elaborates on several important methodological issues related to TDI.

Chapter Five: Discussion

DMM is a recently developed theoretical-psychometric paradigm for educational and psychological research (Dumas et al., 2020). Asymptote estimates from DMM have been used and interpreted as students' capacity or learning potential in different domains. Evidence supporting the validity of DMM capacity scores has also been demonstrated in both empirical and methodological studies (e.g., Dumas & McNeish, 2017; Dumas, McNeish, Clements, & Sarama, 2019; McNeish et al., 2019). Like any formerly established measurement framework, DMM is also undergoing the developmental process, which may involve refinement of methodology, extensions of theories, and further validation in the use and interpretation of produced scores. The present study developed a person-specific DMM trajectory deviance index (TDI) to identify students with aberrant growth trajectories and strengthen the validity of DMM capacity scores.

The Properties and Advantages of the Developed TDI ($\mathrm{TDI}_{\mathrm{U\_Abs}}$)

This research first formulated four different ways to calculate the TDI based on the literature and characteristics of DMM, and $\mathrm{TDI}_{\mathrm{U\_Abs}}$ was determined to be the final index. One might find it surprising that $\mathrm{TDI}_{\mathrm{U\_Abs}}$ performed better than other candidates in almost all aspects of evidence from the current investigation, even in the detection of exponential growth (J-shaped). Theoretically, both weighted TDIs should be more sensitive to detect the growth trajectories that did not level off, but the critical values found at the significance level of .05 could be too conservative. As shown in the second simulation (see Table 2), the two weighted indices presented Type I error rates close to 0 and relatively low detection power. In contrast, the $\text{TDI}_{\text{U\_Abs}}$ presented consistently higher power than the other three TDIs, and meanwhile, it held better Type I error rates (i.e., closer to the ideal range: 025 to .075, Bradley, 1978; Hancock, Lawrence & Nevitt, 2000) and high Rand values across conditions.

Another favorable feature of the $\text{TDI}_{\text{U\_Abs}}$ was the form and consistency of the distribution of this index. TDI measures the magnitude of deviance from student-specific to the model-implied growth curve, which is a quantity that can be used for a wide range of statistical analyses. Given that normality is one of the most fundamental assumptions in statistical analysis (Tabachnick & Fidell, 2014), the normality of $\text{TDI}_{\text{U\_Abs}}$ was thus to be a useful feature for future applications. Moreover, the small standard deviations across replications and the almost identical distributional properties between sample size conditions (see Table 1) both supported the consistency of its distribution.

The $\text{TDI}_{\text{U\_Abs}}$ also shared a common advantage with other DMM methodological advances (e.g., conditional reliability, McNeish & Dumas, 2018). Despite that the conducted simulation was conducted with specific fitting and aberrant growth trajectories and functions, the residual-based TDI, can be applied to different DMM functions, as long as the residual (d_{ti}) has the same definition.

Robustness of DMM Capacity Estimates: Evidence from TDI

In previous work (Dumas & McNeish, 2017; Dumas & McNeish, 2018), DMM capacity scores have demonstrated better consequential validity of measurement than single-timepoint scores. A major piece of evidence was that DMM capacity estimates were less affected by demographic background variables. However, if there was a mediation effect of TDI on this relation, this important finding can be doubted. Specifically, literature shows that different demographics were often associated with mathematical growth and learning (e.g., Clements et al., 2020; Duncan & Magnuson, 2012; Kenney-Benson et al., 2006; Siegler & Booth, 2004), which means students with historical marginalized backgrounds may not follow a typical growth trajectory such as the inverted J-shape. When their capacity scores were generated by the overall model, they might have larger TDI and meanwhile obtain less accurate estimates. It remained a concern whether the weaker association between capacities and demographic background variables was due to the larger trajectory deviances.

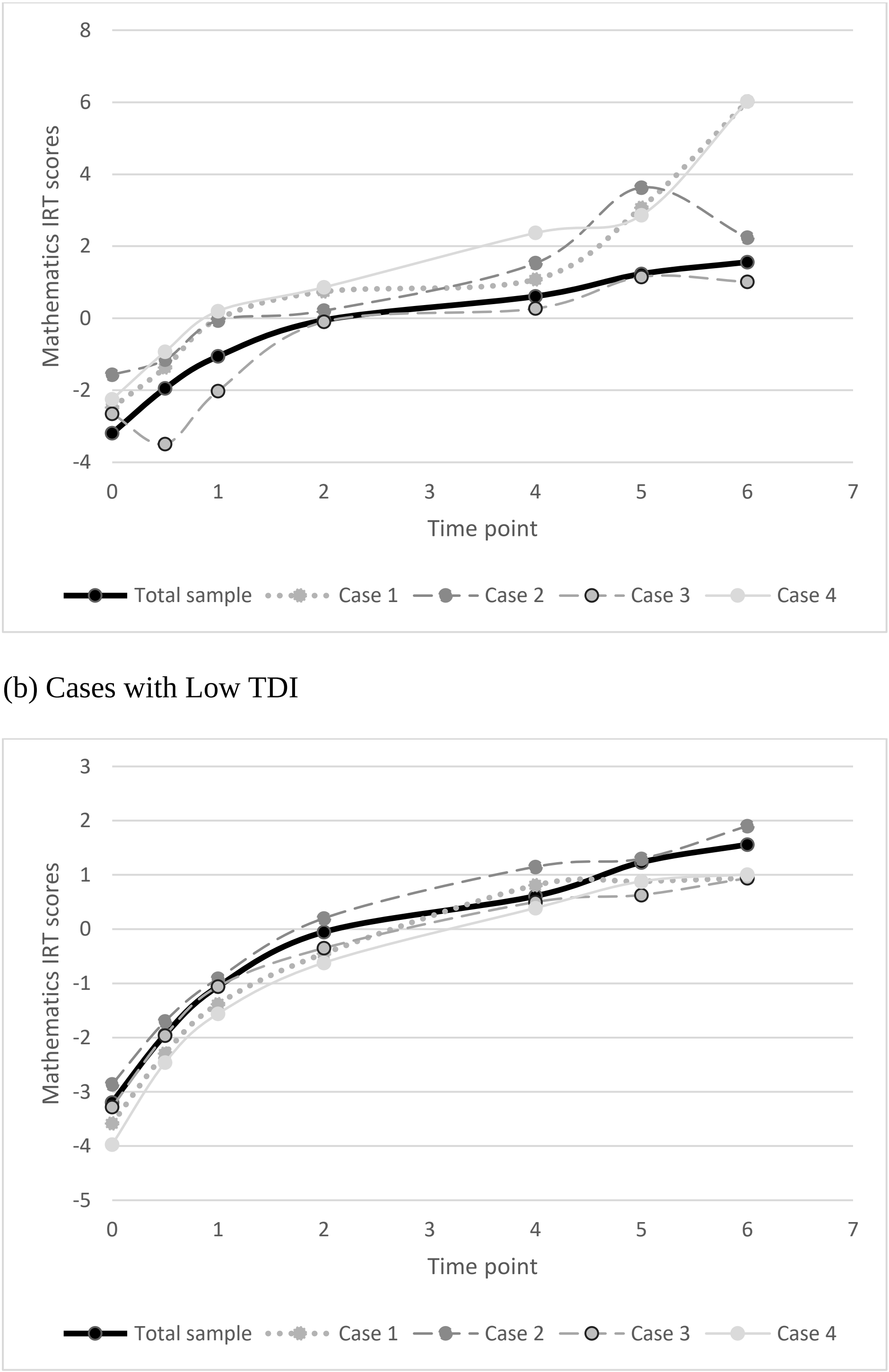
In the second stage of this project, most studied factors (i.e., intervention, gender, age, SES, and special education status) showed non-significant relations to TDI. Although the growth of bilingual students deviated more from their model-implied curves, they did not show substantial deviance in terms of the empirical critical value. There was also no relation between TDI and capacity scores. Therefore, the initially concerned mediation effect of TDI does not exist, and the previous finding regarding DMM consequential validity was not an artifact.

Moving Past ‘One Size Fits All’: Improving DMM Validity by Incorporating TDI

Researchers have been advocating for attention to minority students in educational research (e.g., mathematical learning difficulties for ELL and students with disabilities, Clements & Sarama, 2020; transformative research and evaluation paradigm, Mertens, 2008). This action can usually be echoed in the general research design or goals but neglected at specific steps of implementing research, such as the data analysis.

Appropriate global fit indices of a candidate DMM model often indicated that the model described the data well in general, and these model fit statistics can be easily interpreted as the signal of “good to go.” Unfortunately, such interpretation is not always true, especially in contemporary educational research settings. For example, it has been observed that students’ abilities commonly grow faster in the beginning and eventually level off towards students’ learning capacity (Dumas & McNeish, 2017; Feuerstein, 1979). However, researchers or educational professionals may easily doubt this conclusion by giving examples of students whose growth trajectories do not level off (see Figure 1). A validity threat to DMM then appears because all students received capacity estimates, even if their growth curves were in shapes with no asymptotes (e.g., linear or J-shaped). In other words, the capacity scores these students received are not accurate or meaningful. With the scope of educational equity and social justice, researchers and psychometricians should pursue valid scores for every student rather than only the majority. The current project also made efforts to address this critical goal.

The developed TDI now provides DMM researchers an effective way to recognize student-specific misfitting as well as the violation of the model assumption (e.g., the Michaelis-Menten model assumes learner-specific trajectories tend to level off as time goes forward). For example, Figures 9a and 9b display the growth trajectories of low and high TDI students in the TRIAD data. To ensure the clarity of the figures, each only included the top four or bottom four students along with the averaged growth curve of the total sample. As can be seen in Figure 9a, all four cases with high TDI showed substantial deviance from the model implied growth curve (i.e., inverted J-shaped). Specifically, the growth of cases 1 and 4 started to accelerate in the later timepoints, while cases 3 and 4 experienced learning loss in the later timepoints. The asymptotes for all four cases appeared to be not meaningful, and they also did not show a higher rate of increase in the beginning as described by the Michaelis-Menten function, especially for case 3. In contrast, the four example students with low TDI in Figure 9b presented relatively clear inverted J-shape growth. Both high and low TDI examples further supported the efficacy of the developed TDI. With the TDI approach, researchers are then able to know which students in the data may have received inappropriate learning capacity scores, and follow-up actions could be taken to correct this. Successfully detecting the deviant trajectories has already been an essential enhancement of the DMM validity. Fortunately, additional benefits of incorporating TDI were also found. For example, if researchers choose to remove those deviant cases, they may achieve higher reliability for the rest of the students. After removing 27 (4%) TRIAD students with deviant growth curves, the conditional reliability of DMM increased from .84 to .86. The change in reliability was not obvious because the base reliability was high in this case. Other improvements in psychometric properties may also occur.


(a) Cases with High TDI

(b) Cases with Low TDI

Figure 9. *Examples of Low and High TDI Students in the TRIAD Data*

It should be noted that removing cases may lead to the occurrence of new deviant cases, but when to stop removing cases needs to be carefully determined with both statistical and practical considerations (e.g., the decision tree in the Rasch misfit analysis, Linacre, 2010). Cases deletion in DMM should not be a responsive action to the overall research project but specific to a certain stage of statistical analysis, which means those removed participants from a DMM model can still be studied using other statistical approaches or research paradigms. Moreover, case deletion is not a required or only choice to deal with the aberrant cases detected by the TDI. Alternative approaches like multi-group modeling can also be applied when certain conditions were met (e.g., the sample size is sufficient in each group). As illustrated in the current study, TDI can be treated either as a continuous variable or recoded into a dichotomous variable. Researchers usually have choices to determine the follow-up steps that can strengthen the validity of scores or improve their confidence in using DMM approaches.

Educational Implications

In the U.S., students with certain demographic backgrounds are historically marginalized, and they have been received less necessary instruction and other learning opportunities than their counterparts. Consequently, those students often do not perform well in conventional single-timepoint assessments. Because the conventional scores are commonly used to predict their future performance or as a reference to distribute educational resources, various types of achievement gaps (e.g., racial gaps in the U.S.) are then persistent, and the advance of social justice in the U.S is also hindered. Improving the consequential validity of educational measurement is therefore critical. It has been shown that the capacity scores from DMM were much less impacted by students' demographics than single-timepoint assessment scores (Dumas & McNeish, 2017), which means the newly invented DMM may be a theoretical-psychometric paradigm with better consequentially validity. Students having disadvantaged backgrounds do not show lower learning capacity even though their developed cognitive abilities are still behind. Thus, DMM capacity scores could be a unique and standardizable source for making equitable educational decisions.

In the original DMM framework, all students are assumed to fit the growth model selected based on the overall fit indices, which is a practice of 'One Size Fits All', and the consequential validity cannot be guaranteed for each student. The current study developed the TDI aiming to make DMM research more valid and trustworthy. This index measures the magnitude of deviance from student-specific to the model-implied growth curve. It thus can be used for checking whether the underlying assumption is met for each student. By incorporating the TDI analysis, the consequential validity of educational measurement can also be improved.

Future Directions

The validity of the DMM scores was a central topic across the study. DMM is a relatively new psychometric modeling framework and sufficiently different from CTT and IRT (Dumas & McNeish, 2017; McNeish & Dumas, 2018), but there is no DMM-specific validity framework so far. Either the classic trinity perspective (content, criterion-related, and construct, Cronbach & Meehl, 1955) or the popular duality theory (internal and external, Lissitz & Samuelsen, 2007) of validity can only be partially applied to examine essential topics for DMM. DMM has been gradually receiving more attention since it was invented. DMM researchers and educational psychologists may need a solid DMM validity framework that can be used to evaluate their future work systematically.

Additionally, DMM and the TDI are based on longitudinal data, which is an important feature that differentiates from other measurement frameworks. Incomplete or missing data is always a common and fundamental issue for longitudinal analyses. Though a large amount of research about this topic has been done in the areas of education, medical and general social science (Ibrahim & Molenberghs, 2009), it is still under-researched within DMM. Future studies may also focus on how missingness impacts DMM scores, DMM reliabilities, or the developed TDI in various conditions.

Follow-up studies may also consider establishing a generalizable cut-off and its associated confidence interval. Future studies could try incorporating a standardizer into the current formula and/or conduct simulation analyses focusing on the cutoff value and its CI. However, the author also noticed that the existence of universal cutoff values might not be empirically supported in some cases. For example, Chen et al. (2008) demonstrated that the .05 cutoff for RMSEA, any other value as universal cutoff, or jointly with the CI, were not supported by the simulation results. Thus, the author is not over-optimistic about finding a universal cutoff for TDI at this point, and it is more likely to identify a range of TDI values as the rule of thumb that we can suggest to researchers who will ultimately make the decisions to cut students from the model. Notably, the cutoff or rule of thumb value is only used for making dichotomous evaluation (i.e., fit or not fit), and the current study has also shown its use as a continuous quantity in the real-data application. In other words, TDI does not necessarily need a cutoff value in the research practice it is intended for.

In summary, this dissertation developed a person-specific TDI to measure the magnitude of deviance from student-specific growth curves to the model-implied growth curve, as well as indicate the measurement appropriateness of DMM for each individual. The study simulated and compared the distributional properties and efficacy of four TDI candidates with different formulations. As a result, the $\text{TDI}_{\text{U\_Abs}}$ was the only candidate statistic that demonstrated a consistently normal distribution across replications. It was also the most effective formulation in detecting aberrant growth trajectories base on three evaluation criteria: detection power, Type-I error, and Rand index. Therefore, the $\text{TDI}_{\text{U\_Abs}}$ was chosen to be the final formulation. This study also showed some uses of the developed TDI via an empirical example. The results indicated that the bilingual status of TRIAD students was significantly related to the deviance of growth in early mathematics, but the other examined factors were not. Considerable evidence supported that incorporating TDI into DMM analysis strengthened the validity of score use and interpretation. The current study is an integral part of the developing DMM methodology, because it offers future researchers a quantitative approach to identify the individuals who are not adequately served by the DMM.

References

Abedi, J., & Gándara, P. (2007). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*(4), 36–46. https://doi.org/10.1111/j.1745-3992.2006.00077.x

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*(1), 1–28.

Albers, C. J., Meijer, R. R., & Tendeiro, J. N. (2016). Derivation and applicability of asymptotic results for multiple subtests person-fit statistics. *Applied Psychological Measurement*, *40*(4), 274–288. https://doi.org/10.1177/0146621615622832

Altemeier, L. E., Abbott, R. D., & Berninger, V. W. (n.d.). *NCEN Executive functions for reading and writing in typical*. 20.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aronson, J. (2002). *Improving academic achievement: Impact of psychological factors on education*. Academic Press.

Bailey, D. H., Nguyen, T., Jenkins, J. M., Domina, T., Clements, D. H., & Sarama, J. S. (2016). Fadeout in an early mathematics intervention: Constraining content or preexisting differences? *Developmental Psychology, 52*(9), 1457–1469. https://doi.org/10.1037/dev0000188

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. *Behavioral & Brain Sciences, 11*, 169–232.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*(4), 327–336. https://doi.org/10.1016/j.lindif.2011.01.007

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children’s theory of mind. *Child Development*, *72*(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1-22. https://doi.org/10.1037/h0046743

Chambers, T. V. (2009). The "Receivement Gap": School Tracking Policies and the Fallacy of the "Achievement Gap". *The Journal of Negro Education*, *78*(4), 417-431.

Cheadle, J. E. (2008). Educational investment, family context, and children’s math and reading growth from kindergarten through the third grade. *Sociology of Education*, *81*(1), 1–31. https://doi.org/10.1177/003804070808100101

Cheadle, J. E., & Amato, P. R. (2011). A quantitative assessment of Lareau’s qualitative conclusions about class, race, and parenting. *Journal of Family Issues*, *32*(5), 679–706. https://doi.org/10.1177/0192513X10386305

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological methods & research*, *36*(4), 462–494. https://doi.org/10.1177/0049124108314720

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517–544. https://doi.org/10.1177/0003122419847165

Clements, D. H., Dumas, D., Dong, Y., Banse, H. W., Sarama, J., & Day-Hess, C. A. (2020). Strategy diversity in early mathematics classrooms. *Contemporary Educational Psychology*, *60*, 101834. https://doi.org/10.1016/j.cedpsych.2019.101834

Clements, D. H., & Sarama, J. (2020). *Learning and teaching early math: The learning trajectories approach*. Routledge.

Clements, D. H., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, *36*, 79–90. https://doi.org/10.1016/j.ecresq.2015.12.009

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology*, *28*(4), 457–482. https://doi.org/10.1080/01443410701777272

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*(2), 127–166. https://doi.org/10.5951/jresematheduc.42.2.0127

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812–850. https://doi.org/10.3102/0002831212469270

Cohen, G. L., & Garcia, J. (2008). Identity, belonging, and achievement: A model, interventions, implications. *Current Directions in Psychological Science*, *17*(6), 365–369. https://doi.org/10.1111/j.1467-8721.2008.00607.x

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Crouzevialle, M., & Darnon, C. (2019). On the academic disadvantage of low social class individuals: Pursuing performance goals fosters the emergence of the achievement gap. *Journal of Educational Psychology*, *111*(7), 1261–1272. https://doi.org/10.1037/edu0000349

Cui, Y., & Li, J. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement*, *39*(3), 223–238. https://doi.org/10.1177/0146621614557272

Davis, J., & Martin, D. B. (2008). Racism, assessment, and instructional practices: implications for mathematics teachers of African American students. *Journal of Urban Mathematics Education, 1*(1), 10-34.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, *19*(2), 294–304. https://doi.org/10.1037/0893-3200.19.2.294

Davis-Kean, P. E., & Sexton, H. R. (2009). Race differences in parental influences on child achievement: Multiple pathways to success. *Merrill-Palmer Quarterly*, *55*(3), 285–318. https://doi.org/10.1353/mpq.0.0023

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual Review of Sociology*, *32*(1), 271–297. https://doi.org/10.1146/annurev.soc.32.061604.123127

Dong, Y., Clements, D. H., Sarama, J., Dumas, D., Banse, H. W., & Day-Hess, C. (2020). Mathematics and executive function competencies in the context of interventions: A quantile regression analysis. *The Journal of Experimental Education*, 1–22. https://doi.org/10.1080/00220973.2020.1777070

Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and individual differences, 160*, 1-23. https://doi.org/10.1016/j.paid.2020.109956

Dong, Y., Fan, W., Cheung, F.M., & Li, M. (2020). Development of a short form of the CPAI-A (Form B) with Rasch Analyses. *Journal of Applied Measurement, 21* (4), 515-532.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

DuBois, W. E. B. (2013). *W. E. B. DuBois on sociology and the Black community*. University of Chicago Press. (Original work published 1920)

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*(6), 284–292. https://doi.org/10.3102/0013189X17725747

Dumas, D. G., & McNeish, D. M. (2018). Increasing the consequential validity of reading assessment using dynamic measurement modeling: A comment on Dumas and McNeish (2017). *Educational Researcher*, *47*(9), 612–614. https://doi.org/10.3102/0013189X18797621

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical–psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*(2), 88–105. https://doi.org/10.1080/00461520.2020.1744150

Dumas, D., McNeish, D., Sarama, J., & Clements, D. (2019). Preschool mathematics intervention can significantly improve student learning trajectories through elementary school. *AERA Open*, *5*(4), 233285841987944. https://doi.org/10.1177/2332858419879446

Dumas, D., McNeish, D., Schreiber-Gregory, D., Durning, S. J., & Torre, D. M. (2019). Dynamic measurement in health professions education: rationale, application, and possibilities. *Academic Medicine, 94*(9), 1323–1328. https://doi.org/10.1097/ACM.0000000000002729

Duncan, G. J., & Magnuson, K. (2012). Socioeconomic status and cognitive functioning: Moving from correlation to causation: Socioeconomic status and cognitive functioning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 377–386. https://doi.org/10.1002/wcs.1176

Duncan, G. J., & Magnuson, K. A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *The Future of Children*, *15*(1), 35–54. https://doi.org/10.1353/foc.2005.0004

Dweck, C. S. (2015). Growth. *British Journal of Educational Psychology*, *85*(2), 242–245. https://doi.org/10.1111/bjep.12072

Eccles, J. S. (1986). Gender-roles and women's achievement. *Educational Researcher, 15*(6), 15–19.

Education Commission of the States. (2018). *The nation's report card.* Retrieved from *http://nces.ed.gov/nationsreportcard*

Embretson, S. E. (1987). Toward development of a psychometric approach. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 141–170). Guilford Press.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement*, *26*(1), 88–108. https://doi.org/10.1177/0146621602026001006

Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. University Park Press.

Feuerstein, R., Feuerstein, R. S., Falik, L. H., & Rand, Y. (2002). *The dynamic assessment of cognitive modifiability: The learning propensity assessment device: Theory, instruments and techniques* (Rev. and Exp. ed.). ICELP Publications.

Feuerstein, R., Krasilowsky, D., & Rand, Y. (1974). Innovative educational strategies for the integration of high-risk adolescents in Israel. *The Phi Delta Kappan, 55*(8), 556–558.

Feuerstein, R., Miller, R., Hoffman, M. B., Rand, Y., Mintzker, Y., & Jensen, M. R. (1981). Cognitive modifiability in adolescence: Cognitive structure and the effects of intervention. *The Journal of Special Education*, *15*(2), 269–287. https://doi.org/10.1177/002246698101500213

Feuerstein, R., Rand, Y., Jensen, M. R., Kaniel, S., & Tzuriel, D. (1987). Prerequisites for assessment of learning potential: The LPAD model. In Dynamic assessment: *An interactional approach to evaluating learning potential* (pp. 35–51). Guilford Press.

Flores, Alfinio. (2007). Examining disparities in mathematics education: achievement gap or opportunity gap? *The High School Journal*, *91*(1), 29–42. https://doi.org/10.1353/hsj.2007.0022

Fryer Jr, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American law and economics review, 8*, 249-281. https://doi.org/10.1093/aler/ahl003

Geertz, C. (1963). *Agricultural involution: The processes of ecological change in Indonesia*. University of California Press.

Haberman, C. (1988, April 10). Japan's 'Examination Hell'. *The New York Times*. https://www.nytimes.com/1988/04/10/education/japan-s-examination-hell.html

Hadden, I. R., Easterbrook, M. J., Nieuwenhuis, M., Fox, K. J., & Dolan, P. (2020). Self-affirmation reduces the socioeconomic attainment gap in schools in England. *British Journal of Educational Psychology*, *90*(2), 517–536. https://doi.org/10.1111/bjep.12291

Halpern, D. F. (2014). It’s complicated—in fact, it’s complex: Explaining the gender gap in academic achievement in science and mathematics. *Psychological Science in the Public Interest*, *15*(3), 72–74. https://doi.org/10.1177/1529100614548844

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Lawrence Erlbaum Associates.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling*, *7*(4), 534-556. http://dx.doi.org/10.1207/S15328007SEM0704_2

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745–765. https://doi.org/10.1037/pspp0000075

Henry, D. A., Betancur Cortés, L., & Votruba-Drzal, E. (2020). Black–White achievement gaps differ by family socioeconomic status from early childhood through early adolescence. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000439

Hill, C., Corbett, C., & Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics*. American Association of University Women.

Holloway, S. D., Campbell, E. J., Nagase, A., Kim, S., Suzuki, S., Wang, Q., Iwatate, K., & Baak, S. Y. (2016). Parenting self-efficacy and parental involvement: Mediators or moderators between socioeconomic status and children’s academic competence in Japan and Korea? *Research in Human Development*, *13*(3), 258–272. https://doi.org/10.1080/15427609.2016.1194710

Hooper, F. H., Fitzgerald, J., & Papalia, D. (1971). Piagetian Theory and the Aging Process: Extensions and Speculations. Aging and Human Development, 2(1), 3–20. https://doi.org/10.2190/AG.2.1.b

Horn, J. L. (1967). Intelligence: Why it grows, why it declines. *Transaction, 4*(1), 23-31.

Hui, Y. (2009). The (un)changing world of peasants: Two perspectives. *Journal of Social Issues in Southeast Asia*, *24*(1), 18–31

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *TEST*, *18*(1), 1–43. https://doi.org/10.1007/s11749-009-0138-x

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. J. (2018). Do high-quality kindergarten and first-grade classrooms mitigate preschool fadeout? *Journal of Research on Educational Effectiveness*, *11*(3), 339–374. https://doi.org/10.1080/19345747.2018.1441347

Jennings, J. L., & Bearak, J. M. (2014). “Teaching to the test” in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Jungert, T., Hubbard, K., Dedic, H., & Rosenfield, S. (2019). Systemizing and the gender gap: Examining academic achievement and perseverance in STEM. *European*

Journal of Psychology of Education, *34*(2), 479–500. https://doi.org/10.1007/s10212-018-0390-0

Kane, M. T. (2013). Validating the interpretations and uses of test scores: Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kang, C. Y., Duncan, G. J., Clements, D. H., Sarama, J., & Bailey, D. H. (2019). The roles of transfer of learning and forgetting in the persistence and fadeout of early childhood mathematics interventions. *Journal of Educational Psychology*, *111*(4), 590–603. https://doi.org/10.1037/edu0000297

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, *29*(1), 417–442. https://doi.org/10.1146/annurev.soc.29.010202.100019

Kapur, D., & Perry, E. (2014, May). *Higher education reform in China and India: The role of the state.* Paper presented at Changing Role of State in Asia II: Comparative Perspective, Asia Research Institute, Singapore.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children’s approach to schoolwork. *Developmental Psychology*, *42*(1), 11–26. https://doi.org/10.1037/0012-1649.42.1.11

Kirsch, I., Braun, H., Yamamoto, K., Sum, A (2007). *America's perfect storm three forces changing our nation's future*. Educational Testing Service.

Kuhfeld, M., Gershoff, E., & Paschall, K. (2018). The development of racial/ethnic and socioeconomic achievement gaps during the school years. *Journal of Applied Developmental Psychology*, *57*, 62–73. https://doi.org/10.1016/j.appdev.2018.07.001

Ladson-Billings, G. (2006). From the achievement gap to the education debt: understanding achievement in U.S. schools. *Educational Researcher*, *35*(7), 3–12. https://doi.org/10.3102/0013189X035007003

Lareau, A. (2003). *Unequal childhoods: Class, race, and family life*. University of California Press.

Lee, M. (2003). Korean adolescents' ?examination hell? and their use of free time. *New Directions for Child and Adolescent Development*, *2003*(99), 9–22. https://doi.org/10.1002/cd.63

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*(4), 269–290. https://doi.org/10.2307/1164595

Li, H., & Suen, H. K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education*, *25*(4), 327–346. https://doi.org/10.1080/08957347.2012.714690

Linacre J.M. (2010) When to stop removing items and persons in Rasch misfit analysis? *Rasch Measurement Transactions,* 23(4), 1241

Linacre, J. M. (2020). Winsteps® Rasch measurement computer program User's Guide. Winsteps.com

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational researcher*, *36*(8), 437-448. https://doi.org/10.3102/0013189X07311286

Lohman, D. F. (1999). Minding our p’s and q’s: On finding relationships between learning and intelligence. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 55–76). APA.

Magnuson, K. A., & Duncan, G. J. (2006). The role of family socioeconomic resources in the black–white test score gap among young children. *Developmental Review*, *26*(4), 365–399. https://doi.org/10.1016/j.dr.2006.06.004

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451. https://doi.org/10.3102/1076998614559412

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the cat environment. *Applied Psychological Measurement*, *23*(2), 147–160. https://doi.org/10.1177/01466219922031275

McNeish, D., & Dumas, D. (2020). A seasonal dynamic measurement model for summer learning loss. *Journal of the Royal Statistical Society: Series A*. 1-27. https://doi.org/10.1111/rssa.12634

McNeish, D., & Dumas, D. (2017). Nonlinear growth models as measurement models: A second-order growth curve model for measuring potential. *Multivariate Behavioral Research*, *52*(1), 61–85. https://doi.org/10.1080/00273171.2016.1253451

McNeish, D., & Dumas, D. (2018). Calculating conditional reliability for dynamic measurement model capacity estimates: DMM reliability. *Journal of Educational Measurement*, *55*(4), 614–634. https://doi.org/10.1111/jedm.12195

McNeish, D., Dumas, D. G., & Grimm, K. J. (2019). Estimating new quantities from longitudinal test scores to improve forecasts of future performance. *Multivariate Behavioral Research*, 1–16. https://doi.org/10.1080/00273171.2019.1691484

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Menken, K. (2006). Teaching to the test: How No Child Left Behind impacts language policy, curriculum, and instruction for English language learners. *Bilingual*

Research Journal, *30*(2), 521–546. https://doi.org/10.1080/15235882.2006.10162888

Mertens, D. M. (2008). *Transformative research and evaluation*. Guilford press.

Milner, H. R. (2013). Rethinking achievement gap talk in urban education. *Urban Education*, *48*(1), 3–8. https://doi.org/10.1177/0042085912470417

Molenaar, I.W.,& Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75–106.

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods*, *23*(2), 351–362. https://doi.org/10.1037/met0000132

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960–1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles*, *39*, 21–43.

OECD. (2015). *The ABC of gender equality in education: Aptitude, behavior, confidence*. Retrieved from https://www.oecd-ilibrary.org/education/the-abc-of-gender-equality-in-education_9789264229945-en

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology*, *26*(3), 291–310. https://doi.org/10.1006/ceps.2000.1052

Pattillo-McCoy, M. (1999). *Black picket fences: Privilege and peril among the Black middle class*. University of Chicago Press.

Phillips, K. J. R. (2010). What does “highly qualified” mean for student achievement? evaluating the relationships between teacher quality indicators and at-risk students' mathematics and reading achievement gains in first grade. *The Elementary School Journal*, *110*(4), 464–493. https://doi.org/10.1086/651192

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing* ? *Educational Measurement: Issues and Practice*, *33*(4), 4–12. https://doi.org/10.1111/emip.12045

Plante, I., O'Keefe, P. A., Aronson, J., Fréchette-Simard, C., & Goulet, M. (2019). The interest gap: How gender stereotype endorsement about abilities predicts differences in academic interests. *Social Psychology of Education*, *22*(1), 227–245. https://doi.org/10.1007/s11218-018-9472-8

Popham, W. J. (2001). Teaching to the test. *Educational Leadership*, *58*(6), 16-20.

Potter, D., & Morris, D. S. (2017). Family and schooling experiences in racial/ethnic academic achievement gaps: A cumulative perspective. *Sociological Perspectives*, *60*(1), 132–167. https://doi.org/10.1177/0731121416629989

Potter, D., & Roksa, J. (2013). Accumulating advantages over time: Family experiences and social class inequality in academic achievement. *Social Science Research*, *42*(4), 1018–1032. https://doi.org/10.1016/j.ssresearch.2013.02.005

Quinn, D. M. (2015). Kindergarten Black–White test score gaps: re-examining the roles of socioeconomic status and school quality with new data. *Sociology of Education*, *88*(2), 120–139. https://doi.org/10.1177/0038040715573027

Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school: Trends and predictors. *Educational Researcher*, *44*(6), 336–346. https://doi.org/10.3102/0013189X15598539

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846-850.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*, *47*(5), 284–294. https://doi.org/10.3102/0013189X18762105

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 339–358). Russell Sage Foundation.

Rey, A. (1934). A method for assessing educability. *Archives de Psychologie*, 53, 297–337.

Ribner, A. D., Willoughby, M. T., & Blair, C. B. (2017). Executive function buffers the association between early math and later academic skills. *Frontiers in Psychology*, *8*, 869. https://doi.org/10.3389/fpsyg.2017.00869

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental psychology, 50*(4), 1262-1281.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *Urban Rev, 3*, 16–20. https://doi-org.du.idm.oclc.org/10.1007/BF02322211

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55*(1), 3-38.

Rvachew, S., Thompson, D., & Dey, R. (2020). Can technology help close the gender gap in literacy achievement? Evidence from boys and girls sharing eBooks. *International Journal of Speech-Language Pathology*, *22*(3), 290–301. https://doi.org/10.1080/17549507.2019.1692905

Santos, K. C. P., de la Torre, J., & von Davier, M. (2020). Adjusting person fit index for skewness in cognitive diagnosis modeling. *Journal of Classification*, *37*(2), 399–420. https://doi.org/10.1007/s00357-019-09325-5

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *1*(2), 89–119. https://doi.org/10.1080/19345740801941332

Sarama, J., & Clements, D. H. (2013). Lessons learned in the implementation of the TRIAD scale-up model: Teaching early mathematics with trajectories and technologies. In T. Halle, A. Metz, & I. Martinez-Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 173–191). Paul H Brookes.

SAS (2015). *SAS/STAT® 14.1 User's Guide*. SAS Institute Inc.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Meiji Tokyo.

Scammacca, N., Fall, A.-M., Capin, P., Roberts, G., & Swanson, E. (2020). Examining factors affecting reading and math growth and achievement gaps in grades 1–5: A cohort-sequential longitudinal approach. *Journal of Educational Psychology, 112*(4), 718–734. https://doi.org/10.1037/edu0000400

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*(2), 336–356. https://doi.org/10.1037/0033-295X.115.2.336

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly, 50*(2), 219–232. https://doi.org/10.1002/rrq.92

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development*, *75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x

Sijtsma, K. (1993). Psychometric issues in learning potential assessment. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential* *assessment: Theoretical, methodological and practical issues* (pp. 175–193). Swets & Zeitlinger Publishers.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*(2), 191-207.

Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times: A new person-fit statistic. *Journal of Educational Measurement*, *55*(4), 457–476. https://doi.org/10.1111/jedm.12188

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices an introduction. In M. del Rosario Basterra, E. Trumbull, & G. Solano-Flores (Eds.). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. Routledge.

Spelke, E. S. (2003). Core knowledge. In N. Kanwisher & J. Duncan (Eds.), Attention and Performance: Vol. 20. *Functional neuroimaging of visual cognition* (pp. 29–56). Oxford University Press.

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. *American Psychologist*, *60*(9), 950–958. https://doi.org/10.1037/0003-066X.60.9.950

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811. http://dx.doi.org/10.1037/0022-3514.69.5.797

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379 – 440). Academic Press. doi:10.1016/S00652601(02)80009-0

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential.* Cambridge university press.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., Jukes, M., & Bundy, D. A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence, 30*(2), 141–162. https://doi.org/10.1016/S0160-2896(01)00091-5

Swanson, H. L. (1995b). Effects of dynamic testing on the classification of learning disabilities: The predictive and discriminant validity of the Swanson-Cognitive Processing Test. *Journal of Psychoeducational Assessment, 13*(3), 204–229. https://doi.org/10.1177/073428299501300301

Tabachnick, B. G. & Fidell, L. S. (2014).*Using multivariate statistics*. Pearson.

Telese, J. A. (2012). Middle school mathematics teachers' professional development and student achievement. *The Journal of Educational Research, 105*(2), 102–111. https://doi.org/10.1080/00220671.2010.521209

Thompson, R. A., & Zamboanga, B. L. (2004). Academic aptitude and prior knowledge as predictors of student achievement in introduction to psychology. *Journal of Educational Psychology*, *96*(4), 778–784. https://doi.org/10.1037/0022-0663.96.4.778

Tzuriel, D. (2001). *Dynamic assessment of young children*. Kluwer Academic.

U.S. Census Bureau. (2018). *Income and Poverty in the United States*. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2019/demo/p60-266.pdf

von Stumm, S. (2017). Socioeconomic status amplifies the achievement gap throughout compulsory education independent of intelligence. *Intelligence*, *60*, 57–62. https://doi.org/10.1016/j.intell.2016.11.006

Vygotsky, L. S. (1931/1997). The history of the development of higher mental functions. In R. W. Rieber (Ed.), *The collected works of L S Vygotsky* (Vol 4). Plenum Press.

Vygotsky, L. S. (1962). *Thought and language*. MIT Press. (Original work published 1934).

Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes*. Harvard University Press.

Wilkins, A., & Education Trust Staff. (2006). *Yes we can: Telling truths and dispelling myths about race and education in America*. The Education Trust. Retrieved from http://www.edtrust.org/dc/publication/yes-we-can-tellingtruths-and-dispelling-myths-about-race-and-education-in-america

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*. Mesa Press.