COMBATTING CEILING EFFECTS: MODELING HIGH-ABILITY STUDENT GROWTH USING MULTILEVEL TOBIT REGRESSION

by

Julia Hujar

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Educational Measurement, Research, and Evaluation

Charlotte

2022

Approved by:

_____
Dr. Richard Lambert

_____
Dr. Stella Kim

_____
Dr. Kyle Cox

_____
Dr. Michael Matthews

ABSTRACT

JULIA HUJAR. Combatting Ceiling Effects: Modeling High-Ability Student Growth Using Multilevel Tobit Regression.
(Under the direction of DR. RICHARD LAMBERT)

Pressures associated with accountability testing have resulted in a narrowing of both the curriculum and pedagogy that does not meet the needs of high ability learners. This study proposed that either a different measurement (an above-level computer adaptive assessment) or a different model (Tobit model) should be used to more accurately demonstrate high ability student achievement and growth in order to lessen the pressures on teachers and therefore create an environment better suited for high ability student learning. To answer the research questions under study, a two-part design was used. The first part of the study used an above-level assessment and imposed an artificial ceiling at grade-level with the goal of using Tobit modeling to reproduce uncensored growth estimates using censored data. The second part of the study used naturally censored data with the goal of increasing growth estimates through Tobit modeling. Ultimately, the Tobit models using artificially censored data were able to come close to replicating the uncensored growth estimates under certain conditions. The results indicated that Tobit regression was necessary when examining homogeneous groups of high ability students. Finally, the Tobit regression models were able to increase the growth estimates for high ability students using naturally censored data. The degree to which the models increased, and under which conditions the increases existed are described in detail.

ACKNOWLEDGEMENTS

I would like to extend several special acknowledgments to the people who helped me achieve this degree. First, I would like to thank each member of my dissertation committee for helping me make it to the finish line. Next, I would like to thank my friends and family for the countless hours spent talking to me on the phone, sitting with me at Optimist Hall, and the many texts of encouragement which were all incredibly helpful. Finally, I would like to acknowledge the two organizations that provided data for this project: Northwest Evaluation Association and North Carolina Education Research Data Center.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AG	Academically Gifted
AYP	Adequate Yearly Progress
CAT	Computer Adaptive Test
CCSS	Common Core State Standards
CSEM	Conditional Standard Error of Measurement
CTT	Classical Test Theory
DIF	Differential Item Functioning
EOG	End-of-Grade
ESSA	Every Student Succeeds Act
HLM	Hierarchical Linear Modeling
ICC	Intraclass Correlation Coefficient
IRT	Item Response Theory
ML	Maximum Likelihood
NC	North Carolina
NCE	Normal Curve Equivalent
NCERDC	North Carolina Education Research Data Center
NCLB	No Child Left Behind
NCSCS	North Carolina Standard Course of Study
NWEA	Northwest Evaluation Association
OLS	Ordinary Least Squares
RIT	Rasch Unit

SEM	Standard Error of Measurement
VAM	Value Added Modeling
ZPD	Zone of Proximal Development

CHAPTER 1: INTRODUCTION

Accountability-Driven Assessments

The Elementary and Secondary Education Act, originally passed by President Johnson in 1965, initiated a national focus on the improvement of public education. This act has been reauthorized several times since its original conception, such reauthorizations coming as the No Child Left Behind Act (NCLB; 2001) and the Every Student Succeeds Act (ESSA; 2015). NCLB was passed in an effort to increase achievement in public K-12 schools via increasing accountability pressures on teachers, schools, districts, and states. The language of NCLB focused on schools meeting Adequate Yearly Progress (AYP) goals which were defined as specific percentages of students at grade-level proficiency in reading and mathematics, with proficiency increasing yearly. Intense attention was paid to the proficiency of specific subgroups of students and the achievement gap was thereby quantified (Myers, 2012). NCLB was problematic, however, since states defined proficiency and no credit was given to schools who were able to grow student proficiency levels over the course of the school year if those students still did not meet grade-level proficiency (Ho, 2008).

Although NCLB initially encouraged holding districts, schools, and teachers accountable by examining snapshots of student achievement (percent proficient in a given subject and year), issues were raised about the efficacy of using single time points that did not account for student ability at previous time points (Amrein & Berliner, 2002). ESSA replaced NCLB in 2015 and with this new title came a few changes. States became responsible for creating their own plans for raising student achievement and consequences for not meeting these goals. Additionally, the shift to growth goals in addition to AYP was the solution that intended to address entering achievement of students rather than solely their proficiency status. Although methods of

calculation and status designation vary slightly between states, there are typically three levels of growth which are determined by comparing actual student growth to an a priori expectation (SAS, 2021). These models produce growth composite scores which are then translated into the categories of did not meet, met, or exceeded expected growth. Importantly, in many states, teacher-level growth scores are used as one of the standards by which teachers are evaluated, or in some cases awarded monetary compensation for performance. Based on these composite growth scores and proficiency in mathematics, reading, and science, schools can earn performance grades by weighting growth (20%) in addition to percent proficient (80%; North Carolina Department of Public Instruction, 2017), though formulas differ by state.

Although state-created assessments differ, all states have published technical reports for their accountability-driven standardized summative assessments. Within these technical reports are sections devoted to the process of item and test development; the information that follows came directly from these technical reports (NCDPI, 2006; NCDPI, 2008). For the state of North Carolina, item writers were trained across multiple days on several aspects of item writing and were instructed to create specific percentages of items at varying degrees of difficulty. Specifically, they were informed that easy questions should be answerable by more than 70% of examinees, medium level questions should be answerable by 50-60% of examinees, and hard questions should be answerable by 20-30% of examinees. The item pools needed to consist of 25% easy and hard questions, and 50% medium level questions. Although 25% of the items in the item pool should be considered hard, it is important to note that these items were still created using grade-level content and standards. There are no above-grade items on North Carolina (NC) End-of-Grade (EOG) tests. Items were also rated by writers in terms of Depth of Knowledge levels and Revised Bloom's Taxonomy. Item alignment to standards (both the North Carolina

Standard Course of Study and Common Core Standards) was examined and items were reviewed. After additional rounds of item revision and field testing, item analyses were conducted. These results determined which items to keep, reserve (use sparingly), or delete. Finally, standard setting was completed with groups of trained teachers using the bookmark method. Groups were instructed to set the standard for level three (grade-level proficiency) with the first item that students who had "just barely" met minimum proficiency could be expected to get correct. Grade-level proficiency, or achievement level 3, is considered "sufficient command" of material. An independent alignment study for the 3rd and 7th grade mathematics EOG examined content coverage and performance expectations (similar to Depth of Knowledge levels), and found that the standards used language covering deeper levels of knowledge than the tests required (Smithson, 2015). Ultimately, these tests are reliable and their interpretations are valid for their intended purpose of assigning students to achievement levels. However, the detailed description of the creation of these tests and standard setting procedures (which will be described in the chapters that follow) indicate that grade-level proficiency standards are essentially minimum competency requirements.

Numerous studies have examined the impact of accountability-driven standardized testing on classroom pedagogy. Due to the pressure placed on teachers to get their students to perform well, test preparation activities in classrooms dramatically increased (Au, 2007; Popham, 2001). Several studies determined that these test preparation strategies detracted from high-level instruction (Amrein & Berliner, 2002; Diamond, 2007; Koretz, 2008) and targeted students at middle levels of achievement (Booher-Jennings, 2005; Bulkley, et al., 2010). Blazar and Pollard (2017) conducted surveys and classroom observations to collect data about the rigor of pedagogy when focused on test preparation. Through analyses of both self-reported measures

and coding of lessons, the researchers found that test preparation activities were a significant and negative predictor of ambitious and inquiry-oriented mathematics instruction. These studies will be explored in more detail in chapter two of this dissertation.

Although the transition from NCLB-era snapshots of student proficiency to ESSA-era growth modeling procedures could be considered a more appropriate metric of student achievement across time, the resulting changes to pedagogy and lack of transferability of knowledge from the narrowing of both the curriculum (Berliner, 2011; Diamond, 2007; Farvis & Hay, 2020; Fuller & Ladd, 2013; Rooney, 2015) and pedagogy (Amrein & Berliner, 2002; Resnick, 2010; Welsh et al., 2014) have negative implications for true student learning. Additionally, these tests were designed and validated for the purposes of measuring student achievement and classifying students into achievement levels, not for the evaluation of teachers (American Educational Research Association et al., 2014; North Carolina Department of Public Instruction, 2009). The use of these tests for teacher evaluation and, in some states, monetary incentives is not supported by the test validation process. This problem has been shown to become amplified when teachers have large homogeneous groups of high-ability students, for whom these tests may not capture true ability given research that has demonstrated large numbers of students begin the year above grade-level in both reading and mathematics (Peters et al., 2017).

High-Ability Students

Researchers and practitioners in the field of gifted education refer to students with high intellectual and/or academic abilities in different ways. Among the variety of ways these students may be referred to may include gifted students, advanced learners, high-achievers, or high-ability students. Additionally, these students are defined in different ways across the field of study and

across states and school districts (McBee & Makel, 2019). The National Association for Gifted Children (NAGC) published a position paper in 2019 detailing their definition of giftedness for the purpose of guiding practice. They specifically explained, “students with gifts and talents perform- or have the capability to perform- at higher levels compared to others of the same age, experience, and environment in one or more domains. They require modifications(s) to their educational experience(s) to learn and realize their potential” (p. 1) They do not quantify the degree of performance above peers since this varies by locale (Peters et al., 2019). In this paper I will use the naming convention of high-ability students and will use the definition that includes students who score above the 94th percentile on nationally- or locally-normed standardized achievement tests (the top 5% of students) out of a necessity to quantify the population for the purpose of statistical analyses.

The nature and needs of high-ability students have been thoroughly researched. Succinctly, these students require both social-emotional and academic differentiation that includes increased rigor in the classroom, homogeneous grouping with like-minded peers, and appropriate ways to demonstrate their ability and creativity (Davis et al., 2011; Maker & Schiever, 2005; Plucker & Callahan, 2014a; Renzulli, 1986; Robins et al., 2020). University programs across the country have designed graduate-level training for teachers of high-ability students that cover the nature and needs of high-ability students, their social-emotional needs, and appropriate methods of differentiation for the purposes of increasing academic rigor, engagement, and developing talent (The University of North Carolina at Charlotte, n.d.).

The current federal requirements that include the use of accountability-driven grade-level standardized achievement tests with all students and for the evaluation of all teachers do not take into account the differing needs of high-ability students. The field of gifted education has

explored and discussed the impact of legislation such as NCLB and ESSA on high-ability students. Specifically, they have collectively worried that increased pressures associated with accountability-driven standardized testing would force teachers to focus on test preparation and teaching to the middle rather than focus on high levels of inquiry and pushing students toward excellence (Gallagher, 2004; Gentry, 2006; Kaplan, 2004; Subotnik et al., 2011; Welsh, 2011). Moon et al. (2003) examined a national survey of elementary school teachers and examined qualitative case-studies to determine the impact of standardized assessments on classroom practices and whether state standardized testing programs were a "friend or foe" of gifted education. The researchers found that teachers were using one-size-fits-all practices rather than effective strategies for high-ability students. Although these pedagogical choices are often dubbed "one-size-fits-all," these practices certainly do not fit the needs of high-ability students given the research that shows many of these students are beginning the school year at least an entire grade-level ahead (Peters et al., 2017). Therefore, pressures associated with accountability testing have created environments that do not lead to learning for high-ability students.

Assessment research in gifted education has focused primarily on issues surrounding identification and program evaluation rather than academic progress (Cao et al., 2017). This is likely due to the field's consensus that grade-level standardized achievement tests do not adequately capture the achievement of high-ability students due to the presence of ceiling effects with these tests (McBee, 2010; Subotnik et al., 2011; Warne, 2012; Warne, 2014). Ceiling effects in educational measurement occur when examinees score at or near the maximum obtainable score possible for the test. Additionally, the use of grade-level achievement tests has been demonstrated to be even more problematic when attempting to capture the growth of these students (McCoach et al., 2013; Ryser & Rambo-Hernandez, 2013). Since high-ability students

are not considered a focus subgroup of students that accountability pressures are concentrated around in most states, this issue has largely gone unnoticed or ignored.

Research Problem and Purpose

Ultimately, it seems unlikely that legislation will ever remove the use of grade-level standardized achievement tests for the purposes of gathering student achievement data and for the evaluation of teachers, schools, districts, and states. Unfortunately, the use of these tests with high-ability students is problematic due to the existence of ceiling effects for this population. Therefore, if the tests themselves cannot adequately measure the achievement of high-ability students, their growth cannot be accurately modeled, and their teachers cannot be appropriately evaluated.

Potential solutions to this problem include either implementation of a better measure of academic performance or a better model of student growth when using a flawed measure. Warne (2012) conducted a comprehensive literature review on the topic of above grade-level testing and found that five reasons were typically presented in favor of this method of assessment: (a) raising the test ceiling, (b) increasing score variability and discrimination, (c) improving reliability, (d) sound interpretations of results, and (e) reducing regression towards the mean. Although the strength of evidence surrounding above-level testing varied, this type of test may also be better able to capture student growth and therefore provide a more accurate evaluation of teachers of the gifted.

Although Maximum Likelihood (ML) estimation is typically considered the gold standard for estimation, there are still issues in estimation when the measure used has ceiling effects. An in-depth explanation of these and related measurement issues and modeling problems will be detailed in chapter two.

Tobit modeling (Tobin, 1958) uses a modified version of ML estimation that has been used abundantly in econometrics (e.g., Amore & Murtinu, 2019; Barros et al., 2018; McDonald & Moffit, 1980) with data that have ceiling effects. Only once has it been suggested for use with educational data (McBee, 2010). McBee demonstrated with simulated data that Tobit modeling produced simple linear regression estimates for a censored sample that were closest to those in the uncensored sample. Despite this success, Tobit modeling has not yet been used in an applied educational setting or with multilevel growth modeling.

The purpose of this research is two-fold: (1) to demonstrate the true growth of high-ability students using an above-level assessment, and (2) to use Tobit modeling with a grade-level (i.e., censored) assessment to attempt to replicate true growth for high-ability students.

Delimitations

One delimitation worthy of discussion is the lack of a definition for high-ability students across the field of gifted education. However, the topic of high-ability is nuanced, and therefore the lack of a standard definition is necessary in order to encompass the diverse gifts and talents that students possess. The choice to use the 95th percentile (on a nationally- or locally-normed assessment) as a cutoff for high-ability students was based on the necessity to study students who score on the upper range of assessments in order to assess the impacts of ceiling effects. A related delimitation is the use of the dichotomous "gifted" variable from the North Carolina Education Research Data Center (NCERDC) EOG data set. Although these students are defined as Academically Gifted (AG) in either reading or math, since the data have been de-identified, there is no way of knowing the specific identification criteria by which these students were labeled, and therefore results of analyses using this particular variable should be interpreted with this in mind. However, the broad state definition defines AG students in the following way:

Academically or intellectually gifted (AG) students perform or show the potential to perform at a substantially high levels of accomplishment when compared with others of their age, experiences or environment. Academically or intellectually gifted students exhibit high performance capability in intellectual areas, specific academic fields, or in both the intellectual areas and specific academic fields. (Article 9B, N.C.G.S. 115C-150.5; NC DPI, 2015).

This study specifically focuses on students in grades 3-8. Therefore, any implications of this research should be limited to such grades. This focus was determined due to testing in grades 9-12 being much more varied by subject and difficulty. Additionally, the tests used in NC as EOG tests in reading and mathematics for grades 3-8 have published technical manuals with important information regarding norming procedures, item writing, test construction, and psychometric properties of the assessments

Another important caveat that must be considered while reading this dissertation is that the problems expressed here regarding measurement of student achievement and modeling of student growth are only problematic for the specific group of aforementioned high-ability students, for whom grade-level assessments have ceiling effects. These tests have been designed and validated for the purposes of measuring student achievement in order to classify students working on or near grade level into broad levels, with special attention given to students of average ability and students with disabilities. Although the use of these tests for teacher evaluation was not an intended application, this usage is less problematic for students whose ability and growth can be adequately measured and modeled.

Definitions

Above-level testing: The use of an assessment that includes material from ability levels higher than that of the tested students' ability level for the purposes of allowing students to demonstrate their ability without the presence of ceiling effects.

Ceiling effect: As defined by the APA Dictionary, "a situation in which the majority of values obtained for a variable approach the upper limit of the scale used in its measurement" (APA Dictionary of Psychology, n.d.).

Censored data: As defined by the APA Dictionary, "a set of data in which some values are unknown because they are not observed or because they fall below the minimum or above the maximum value that can be measured by the scale used" (APA Dictionary of Psychology, n.d.).

Educator Value Added Assessment System (EVAAS): A proprietary modeling suite that is contracted by the state of North Carolina (and others) for the calculation of value-added models for student growth as well as teacher, school, and district accountability (SAS, 2021).

End-Of-Grade (EOG): The assessment used by the state of North Carolina to assess students in grades 3-8 in mathematics and reading as well as science in grades 5 and 8.

Every Student Succeeds Act (ESSA): a 2015 reauthorization of the Elementary and Secondary Education Act of 1965 that continued accountability pressures from NCLB but also added in growth measures as a relevant portion of the formulas used to calculate school performance grades. Additionally, this legislation required individual states to create their own plans to increase student proficiency.

Growth modeling: In the context of this study, this term can be defined as an umbrella covering modeling options that account for student initial proficiency when examining end-of-year proficiency.

Longitudinal multilevel modeling: Linear regression that accounts for nested data structure with measurements across time nested within students, students nested within teachers or classes, and teachers/classes nested within schools/districts.

Maximum Likelihood Estimation (MLE): As defined by the APA Dictionary, "a statistical technique in which the set of possible values for the parameters of a distribution is estimated based on the most probable sample of observations that one might have obtained from that population" (APA Dictionary of Psychology, n.d.).

Measure of Academic Progress (MAP): An above-level assessment created by the Northwest Evaluation Association (NWEA) that is administered three times per year for the purpose of tracking student growth throughout the year and across years. This assessment is now referred to as the MAP Growth assessment.

Multilevel Tobit Regression: As defined by Springer, "a family of statistical regression models that describe the relationship between censored or truncated continuous dependent variables and some independent variables" (Springer Link Encyclopedia of Quality of Life and Well-Being Research, n.d.).

No Child Left Behind Act (NCLB): A 2001 reauthorization of the Elementary and Secondary Education Act of 1965 that increased accountability pressures in an effort to increase proficiency in mathematics and reading across the United States via the use of language such as "Adequate Yearly Progress" (AYP).

Normal curve equivalent (NCE): A normalized standard score with a range of 1-99. NCEs have a mean of 50 and standard deviation of 21.06 (Crocker & Algina, 1986).

Stata: A statistical modeling software (https://www.stata.com/)

Value added modeling (VAM): An umbrella of modeling options that attempt to quantify the value to a students' education that is added by a particular teacher, school, or district. (SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, n.d.)

Vertical Scaling: As defined by SAGE Encyclopedia, "a special form of linking, which aims at adjusting score differences on tests that differ in content and/or difficulty...Vertical scaling is intended to establish the concordance relationship between scores on tests measuring educational achievement or aptitude at different academic grades" (SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, n.d.).

CHAPTER 2: LITERATURE REVIEW

In this review of literature, I will explore the extant research on the topics of grade-level assessments, growth and value-added modeling, and potential solutions to the issues presented by the aforementioned measures and models. First, the review will cover the use of accountability-driven grade-level standardized achievement tests and associated problems with their use with high-ability students. Next, the review will explore various growth and value-added models and their uses and problems in both general and gifted education. Finally, the two proposed potential solutions (i.e., potentially better measure or potentially better model) will be discussed in detail. Ultimately, the review of literature will conclude with specific research questions for this study.

Grade-Level Assessments

Learning Environment

The implementation of accountability-driven grade-level standardized achievement assessments has impacted both educational policy and practice. States and districts have been found to make questionably ethical structural changes in response to such accountability pressures (Delisle, 2014; Sykes, 1995). Increased pressures to meet adequate yearly progress (AYP) have forced schools to make decisions about staffing, teacher assignments, and resource allocation. Fuller and Ladd (2013) examined fifteen years of teacher data from the state of North Carolina to determine the allocation of teachers based on teacher licensure test scores and degree data weighted by value-added scores (i.e., a measure of teacher quality) across elementary grade levels. The researchers found that school principals strategically placed stronger teachers in the upper grades where accountability pressures were higher and that teachers with high licensure test scores were more likely to be moved from a lower to higher grade-level after the

introduction of NCLB. An older study examined responses to the pre-NCLB Texas accountability system (Booher-Jennings, 2005). The results of this study indicated that schools engaged in "educational triage" by allocating more resources to students at the cusp of passing and to those groups of students who heavily impacted the school's accountability rating. This study posited that this likely occurred due to the equating of good teaching with high test scores. Similarly, Farkas and Duffett (2008) conducted a large nationally representative survey to explore teacher opinions related to student ability levels. The researchers found that teachers allocated their time primarily to students below grade level (63%), followed by average students (13%) and finally high ability students (7%). Additionally, they found that 77% of teachers believed that the needs of high ability students were not a top priority.

In addition to certain questionably ethical structural changes as responses to accountability pressures, schools and districts have also narrowed the curriculum that students are taught in order to emphasize tested subjects (i.e., reading and mathematics; Diamond, 2007; Farvis & Hay, 2020; Rooney, 2015). Xie (2013) explained that curriculum narrowing is "characterized by using test materials, following a test-based curriculum, using similar or identical test-items, or focusing exactly on what the test measures" (p. 198). Berliner (2011) documented an average increase of 141 minutes per week in English Language Arts and 89 minutes per week in mathematics. These increases, as explained by Berliner, were made possible by decreasing time spent in social studies, science, physical education, recess, art, and music each by approximately 40-76 minutes per week. The field of study surrounding the narrowing of the curriculum has drawn heavily on teacher voices. Using a combination of teacher observations and a survey of teacher-reported changes made in response to high-stakes testing policies, Diamond (2007) examined data from the Distributed Leadership Project in Chicago and found

that 42-46% of teachers reported that their pedagogy was influenced by standards and/or testing and that teachers linked the influence of high-stakes testing with changes in instructional content. The narrowing of the curriculum has resulted in increased teacher and administrator stress (Farvis & Hay, 2020) and teacher dissatisfaction (Crocco & Costigan, 2007; Rooney, 2015). Crocco & Costigan (2007) interviewed new teachers and found that in response to narrowed curriculum, specifically in the form of scripted lessons, these teachers found their "personal and professional identity thwarted, creativity and autonomy undermined, and ability to forge relationships with students diminished" (p. 512). Rooney (2015) conducted an ethnographic study of teachers' experiences with high-stakes testing accountability pressures and found that instructional mandates resulted in a decrease in teachers' professional discretion and ultimately teachers were unable to find enjoyment in their work. Farvis and Hay (2020) surveyed and interviewed educational consultants in New York schools and determined that high-stakes testing was linked with reduced teacher control in instructional planning, narrowing of the curriculum, less teacher collaboration, and an increase in test preparation strategies.

This increase in test preparation strategies ultimately represents a narrowing of pedagogy in addition to the narrowing of the curriculum (Koretz & Hamilton, 2006; McNeill, 2002; Resnick, 2010; Shepard, 2002; Welsh et al., 2014). Blazar & Pollard (2017) drew from a nationally representative sample of teacher surveys and classroom observations to conclude that test preparation strategies were prevalent in classrooms across the United States. Hujar (2021) interviewed teachers about their perceptions of standardized achievement assessment and all teachers admitted to having changed their preferred pedagogy to include test preparation strategies as a result of administrative pressures to focus on test scores. Resnick (2010) explained that in one district she had studied intensively, she found that "elementary students stopped

reading and discussing grade-level-appropriate books in February and instead spent time digesting brief passages, accompanied by multiple-choice test items that mimic the ones that appear on the state tests" (p. 185) until the end of the school year when testing began.

Despite the increased attention to test-preparation activities, there has not been a subsequent increase in general student achievement. Welsh et al. (2014) examined the efficacy of test preparation activities in 32 third- and fifth-grade classrooms. The researchers concluded based on the results of their hierarchical linear modeling that "instruction on tested objectives using items like those presented on the state test, decontextualized practice, and teaching test-taking skills offered no student achievement benefit relative to general instruction on the state standards" (p. 98). Amrein and Berliner (2002) examined eighteen states with severe consequences associated with their testing programs to determine if learning from domains covered by the high-stakes tests transferred to performance on tests covering similar domains. The researchers found that high stakes testing performance did not translate to an increase in student performance on the alternate measure of student ability. They explained that this result was likely due to test preparation strategies impacting the high-stakes test results more than the alternate measure. These results also highlight an issue raised by Xie (2013) regarding the validity of any increases in achievement test scores when test-preparation strategies are used. Specifically, Xie noted that "if test-takers focus on the narrow range of content and skills that a test samples thereby improving test scores, the inflated scores are unable to represent a corresponding increase on the domain of interest" (p. 198). Ultimately, though test preparation strategies could impact scores on the specific test students have been trained to take (despite Welsh and colleague's evidence that it did not for their particular sample), these test preparation

strategies are not contributing to a true increase in student learning and test score increases (if any) may be artificial.

The narrowed curriculum and pedagogy derived from accountability-driven standardized achievement assessments have resulted in a decrease in the required cognitive demand for all students. In addition to their conclusions regarding an increase in test preparation strategies in classrooms across the country, Blazar and Pollard (2017) also concluded that these test preparation activities were predictive of lower-quality and less ambitious mathematics instruction. Moon et al. (2003) conducted a national survey of elementary school teachers as well as qualitative case studies and found that teachers were not likely to be engaging in effective classroom practices, but opting for "one-size-fits-all" practices which the authors characterized as lacking rigor. Due to budgetary constraints, Berliner (2011) explained, "the items used to assess students are quite often multiple-choice, convergent, machine-scorable items, the cheapest items to produce for mass testing" (p. 296) resulting in low cognitive demands required of students while taking these tests. Specifically, "higher order thinking is sacrificed when high-stakes multiple choice testing puts pressure on teachers" (Berliner, 2011, p. 296). Resnick (2010) argued that the current accountability-driven grade-level standardized achievement tests have pushed the nation back towards the minimum competency movement of the 1970s (Jaeger & Tittle, 1980). Jerald (2006) reported that score differences between third grade students were mostly related to their ability to efficiently and fluently decode words, whereas by tenth grade, score variation was more related to vocabulary and comprehension skills. The narrow focus and low cognitive rigor of these exams led Berliner (2011) to posit that it is possible for students to perform poorly on standardized reading and math tests in later grades where the focus is more on

comprehension and reasoning rather than an emphasis on simple decoding and algorithms on which their earlier years of schooling focused.

Although all students are impacted by the narrowing of the curriculum, pedagogy, and decreased cognitive demands associated with accountability-driven assessments, high-ability students stand to suffer the most from teachers who have shifted their focus to test-preparation strategies targeting average- and low-achieving students in an effort to raise their test scores. Eminent scholars in gifted education have published opinion pieces expressing their concerns about accountability-driven assessment systems' impact on the education of high-ability students (Gallagher, 2004; Gentry, 2006; Kaplan, 2004). Hujar (2021) found in her interviews with teachers of high-ability students that these teachers had cognitive dissonance from knowing that test preparation did not meet the needs of their high-ability students, but also feeling pressure from administration to engage in these pedagogical choices in order to obtain high test and student growth scores. The researcher also determined that these teachers would not have known that test preparation strategies did such a disservice to their high-ability students had they not been participants in graduate-level certification programs in gifted education. This is hugely problematic given the lack of training general education teachers receive on the topic of high-ability student education (Farkas & Duffett, 2008).

To reiterate, high-ability students require both social-emotional and academic differentiation that includes increased rigor in the classroom, grouping with like-minded peers, and appropriate ways to demonstrate their ability and creativity (Davis et al., 2011; Maker & Schiever, 2005; Plucker & Callahan, 2014a; Renzulli, 1986; Robins et al., 2020). Given these needs, test preparation strategies that attempt to prepare students to take minimum competency tests do not meet this need. Peters and colleagues (2017) compared proficiency scores on grade-

level test data from Wisconsin, California, and Texas to the Measure of Academic Progress (MAP) test with a high measurement ceiling to determine that approximately 14-37% of students scored at least an entire year above grade-level in mathematics and 20-49% in English language arts. It has been well researched and demonstrated that students learn when they are met with content that is in their Zone of Proximal Development (ZPD; Vygotsky, 1978); therefore, if high-ability students are presented with low cognitive demand test preparation activities for a test whose mastery they could have demonstrated an entire school year prior, these students are far from their ZPD and very little if any learning will occur. McBee and colleagues (2018) conducted a simulation study and determined that students whose needs were most aligned with classroom content (school curriculum overlapped their ZPD) were likely to grow the most. Ultimately, the pressures associated with accountability-driven grade-level standardized achievement tests have created an environment not fit for learning for high-ability students.

Measurement Issues

In addition to the problems with the learning environment caused by the pressures associated with accountability testing, the grade-level standardized achievement tests used for accountability purposes are riddled with measurement issues when attempting to capture high-ability student achievement (Cross & Cross, 2010; Kieffer et al., 2010; Kline, 2010; Lohman & Korb, 2006; McBee, 2010; Olszewski-Kubilius, 2010; Plucker & Callahan, 2014b; Rambo-Hernandez & Warne, 2015; Subotnik et al., 2011). To demonstrate this point, theoretical true student ability will be examined through both the lens of Classical Test Theory (CTT) and Item Response Theory (IRT). The information that follows in this paragraph is a summary of information obtained from Crocker & Algina (2006). The point of any test is to measure, as accurately as possible, a student's true score on a particular construct; in other words, the goal is

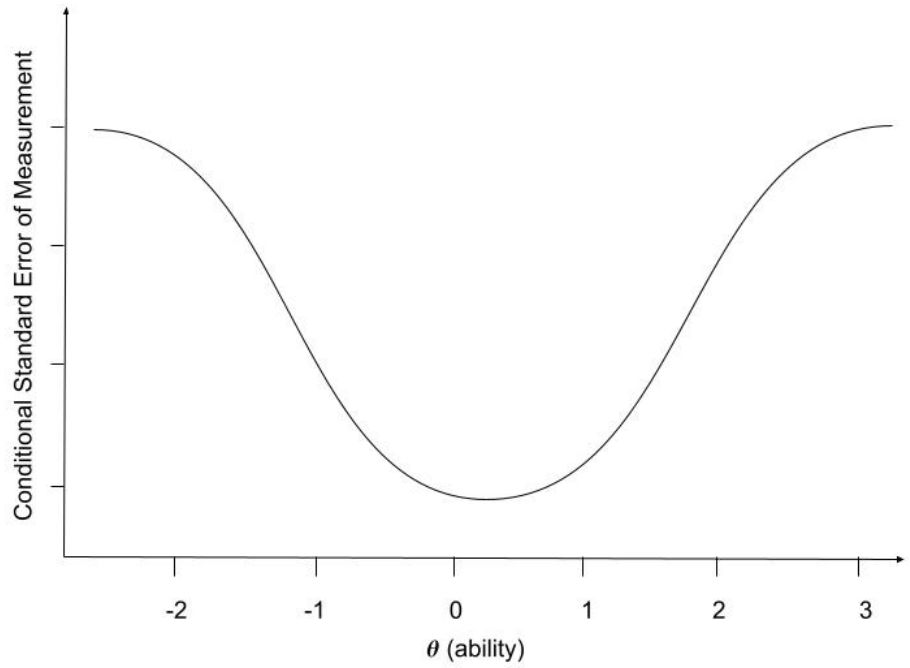
to determine their true ability level. Theoretically, true scores would be obtained by administering the same assessment an infinite amount of times or an assessment with an infinite number of items. However, true scores are not directly observable and therefore the observed score that is obtained from testing contains some degree of measurement error. A simple way of expressing this concept through the framework of CTT is the following representative equation:

$$T = X + E \quad (1)$$

Where T represents an examinee's true score, X represents their observed score, and E represents measurement error. Although measurement error itself is not directly quantifiable, the standard deviation of the distribution of measurement error can be thought of as the level of variability in observed score distributions. This can be further expressed on an individual level by representing the standard deviations of each examinee's personal distributions of observed scores around their true score (Conditional Standard Error of Measurement; CSEM). When using scaled or normative scores (typically the way standardized test scores are reported), CSEM is highest at the extremes of score distributions (Lohman & Korb, 2006). This means that for students at the lowest and highest test scores, their true score is hidden by more measurement error than it is for students near the middle of the distribution (Lohman & Korb, 2006; Welsh, 2011; see Figure 1 for a general depiction of CSEM). This is due to a range of factors, one being that the items on the test are either too difficult or too easy for them. In minimum competency testing situations such as those used for accountability measures, most items on these tests are too easy for high ability students and therefore their observed scores are inaccurate reflections of their true score.

Figure 1

General Depiction of Conditional Standard Error of Measurement

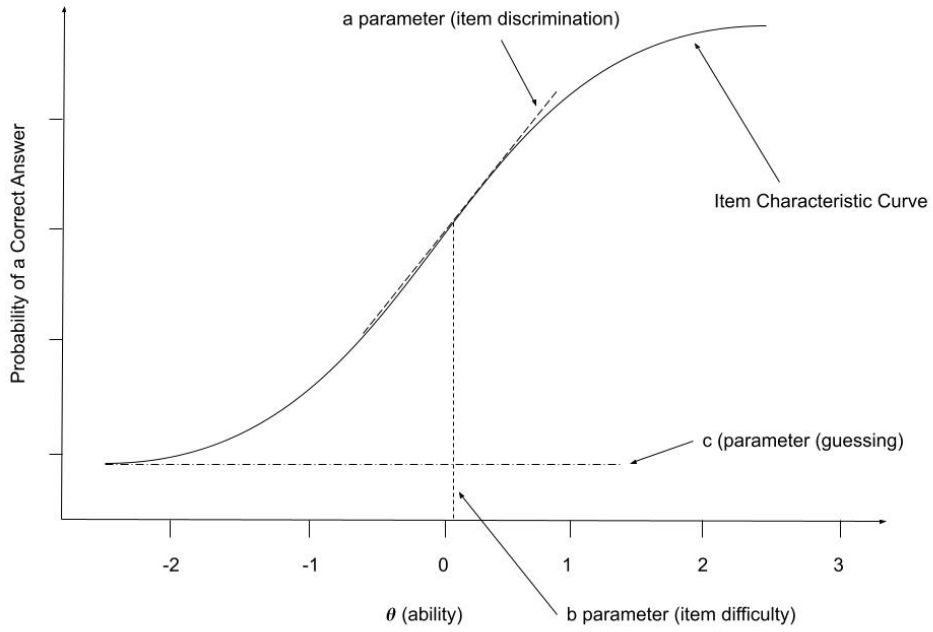


In addition to CTT, IRT can also provide insight into why these minimum competency tests do not adequately measure high-ability student achievement. The information that follows in the next two paragraphs is a summary of information obtained from Lord (1980). At its core, IRT models performance at an item level through item response functions (or Item Characteristic Curves). These item characteristic curves provide the relationship between the probability of correctly answering an item and the ability level of the theoretical examinee (See Figure 2 for an example Item Characteristic Curve). As an individual's ability level increases, so should the probability of them answering an item correctly. Item response functions are typically created based on three parameters (for single-answer, multiple choice items; though some models use two or only one parameter). The *a* parameter is called the item discrimination parameter and determines the rate at which the probability of correctly answering an item changes given differing levels of ability; this parameter is essentially the slope of the curve, with steeper slopes

indicating better discrimination between individuals. The *b* parameter is the item difficulty parameter; this value is the location along the ability scale where respondents have a 50% probability of correctly answering the item. Curves for items with higher difficulty will be located further to the right on the ability scale than easier items. Finally, the *c* parameter (not used in all models) is the guessing parameter, or the probability of guessing the correct answer choice. The item response functions can be used to determine an item information function. The amount of information given by an item varies by ability level. On a basic level, these item information functions can be summed to create test information functions. Information provided by items and tests is determined by both the item difficulty parameter and the item discrimination parameter. Items that have both a high *b* parameter and *a* parameter provide the most information. In practice, tests typically provide less information at higher ability levels because most grade-level tests are designed to have fewer items at a high difficulty level (e.g., North Carolina Department of Public Instruction, 2009).

Figure 2

General Depiction of an Item Characteristic Curve



Tests can be compared to one another by creating a ratio of their test information functions to determine the relative efficiency of one test over another. Similar to item information functions, relative efficiency also varies according to ability levels. Imagine two tests measuring the same construct, one with items spanning a large range of ability levels and one with items whose difficulty is concentrated around the mean ability level; the test with the more varied item difficulties would be more efficient at the extremes of ability levels. Essentially, this test would provide more information about examinees at the extremes of ability levels because of the larger range of item difficulties. However, minimum competency exams like those used for accountability purposes target students of average ability level by concentrating item difficulty around expected average ability levels; therefore, these exams provide less information about examinees with extremely high or low ability levels (Welsh,

2011). Students at the low end of the ability distribution are legally protected by the Individuals with Disabilities Education Act (IDEA; 2004) and are given alternate assessments or testing accommodations in order to gather more accurate information about their ability. No such legal provisions exist for students with high abilities.

Accountability-driven assessments are written to be appropriately challenging and yield the maximum information for the middle 90% of students, though in practice maximum information is potentially obtained for a smaller subset of students. Therefore, the focus of these assessments is grade-level material. This means that since many students are beginning the school year an entire year above-grade level (Peters et al., 2017), these high-ability students could presumably meet proficiency or even mastery of their current grade level on the day that school begins. Thus, by the end of the school year their mastery of the content is likely to be nearing the maximum obtainable score on these minimum competency assessments (Subotnik et al., 2011). This phenomenon of measurement inadequacy is called a ceiling effect (Rambo-Hernandez & Warne, 2015). Since high-ability students are scoring at or near the ceiling of these tests, their true ability likely lies beyond the scope of measurement of the given assessment. Ultimately, grade-level tests used in minimum competency environments (i.e., with low ceilings) do not capture the true achievement of high-ability students (Callahan, 2009; McBee, 2010; Plucker & Callahan, 2014b; Subotnik et al., 2011; Rambo-Hernandez & Warne, 2015; Ryser & Rambo-Hernandez, 2013; Welsh, 2011)

Growth and Value-Added Modeling

Growth modeling, in the context of this study, is the statistical process used to examine change in student ability over time. True statistical growth modeling requires (a) at least three observations, (b) accurate measurement of time, and (c) assessment scores that are

psychometrically sound and comparable over time (McCoach et al., 2013). This can be done via one of two frameworks: Structural equation modeling or hierarchical linear (or multilevel) modeling (HLM). McNeish and Matta (2018) detailed the differences between structural equation modeling (which the authors refer to as the latent-curve approach) and HLM (which the authors refer to as the mixed effect approach) styles of growth modeling.

Mixed effect modeling is done within the framework of regression and allows researchers to include both fixed and random effects when modeling individual growth trajectories. Fixed effects in the context of growth modeling demonstrate the average initial starting score for all students, holding any covariates equal, and the average amount of growth over a unit of time. Random effects display the degree to which the estimates (both initial status and change over time) differ from the fixed effects at various levels of the model (student-level, teacher-level, school-level, etc.). The general form for this type of model can be written as

$$\begin{aligned} Y_{ti} &= \pi_{0i} + \pi_{1i} a_{ti} + e_{ti} \\ \pi_{0i} &= \beta_{00} + \sum_{q=1}^{Q_0} \beta_{0q} X_{qi} + r_{0i} \\ \pi_{1i} &= \beta_{10} + \textstyle\sum_{q=1}^{Q_1} \beta_{1q} X_{qi} + r_{1i} \end{aligned} \tag{2}$$

where Y_{ti} is the observed achievement at time *t* for individual *I*, π_{0i} is the intercept (the achievement of the individual at $a_{ti} = 0$), π_{1i} is the growth rate for person i over time , a_{ti} is the measure of time for individual *i* at time *t*, e_{ti} is the level 1 error, β_{00} represents the average achievement across all individuals, β_{10} represents the average growth rate across all individuals, β_{0q} and β_{1q} represent the fixed effect of X_{qi} on the growth parameters (π_{0i} and π_{1i}, respectively), X_{qi} is an individual-level covariate, and r_{0i} and r_{1i} are the level two random effects (Raudenbush & Bryk, 2002, p. 163).

In contrast, the latent-curve approach is conducted within a general structural equation modeling framework as a confirmatory factor analysis model which yields growth estimates (McNeish & Matta, 2018). Although both frameworks allow for the modeling of overall mean trajectories and individual-level deviations, this approach requires random effects as latent variables in a confirmatory factor analysis rather than regression coefficients. Mathematically, the mixed effect model and the latent curve model are equivalent (McNeish & Mata, 2018). McNeish and Matta (2018) argued that the mixed effects or regression framework is usually preferable for straightforward models, those with time-unstructured data, or models with several levels of nesting. Therefore, for the purposes of the present study, mixed effects modeling is preferable over latent curve modeling due to its ability to more easily create mixed-effects models that account for the nested structure of educational data as well as the time-unstructured nature of the particular data being used.

Castellano and Ho (2013) identified three primary interpretations of growth models: description, prediction, and value-added. Ryser and Rambo-Hernandez (2013) summarized these interpretations, “Growth description provides a growth metric about the magnitude of growth for an individual or group. Growth prediction provides information about the future scores of students given current and past achievements. Value-added provides information about what causes growth, for example particular educators and schools.” (p. 19). While the modeling of growth description and prediction are for the purposes of monitoring student achievement and progress in order to inform instructional and placement decisions, value-added models are primarily used as a method for isolating teacher effects and thereby evaluating teacher quality. Policy makers as well as local education agencies have become more interested in teacher quality

as accountability pressures have increased. The assumption underlying this interest is that better teachers should be able to produce greater gains towards proficiency for their students.

There are several different ways to model student growth, the most basic form being single-wave models. These models examine changes in academic achievement in a particular subject based on two data points (Anderman et al., 2015). Since these types of models only use two time points, these models should be considered quasi-growth models rather than true growth models (McCoach et al., 2013). At its most basic form, the student gain score model determines a student's change in achievement by calculating the difference between two test scores (Anderman et al., 2015). To determine the value added by a particular teacher, average gains between teachers are compared across a school or district. Although simple differences are unbiased estimates of true gain (Anderman et al., 2015; Ragosa et al., 1982; Zumbo, 1999), the shape of the trajectory cannot be modeled, and therefore their use is limited. Additionally, students with missing data are not included in the analyses and previous achievement level is not accounted for. Finally, this type of value added model (VAM) can only be used to determine the teachers whose students show the largest gains within a set unit of comparison (e.g., grade level, school, district). Even using this method to compare teachers within one school is problematic since teachers are often assigned very different populations of students.

Another type of single-wave model is the covariate-adjusted gain model. For this type of model, growth is estimated while adjusting for relevant student- and school-level covariates (e.g., demographics). Although these types of models are slightly better than the simple gain score model, since schools lack random assignment, covariates may not be able to adequately adjust for selection bias due to issues related to resource allocation (Anderman et al., 2015). This type of model still only uses two time points of data despite McCoach and colleagues' (2013)

demonstration that best practices for growth modeling should include more than two time points. However, a potential covariate that could be included in the model is previous achievement; if the model includes previous achievement, it could be considered a true growth model and HLM or structural equation modeling techniques could be used.

A final type of single-wave models are the student percentile gain models. These models use normal curve equivalent scores (NCEs) to compare student location within a group from year-to-year (Anderman et al., 2015). Using NCEs allows states to use assessments that are not vertically scaled. These types of models are plagued by the same issues as the other single-wave models since students with missing data are not included and application is limited.

The ideal way to model student growth for description and prediction is using multilevel growth models that utilize at least three time points of data (McCoach et al., 2013). These growth models require accurate measures of time that has elapsed between testing events. These can be time-structured (testing occurring at the same regular intervals for all students) or time-unstructured (testing occurring at different intervals) but must be measured accurately. Additionally, the assessment scores must be both reliable and valid for the purpose of tracking student achievement over time. Vertically scaled assessments are also necessary in order to compare student performance over time along an equal-interval scale. As previously stated, if these conditions are met, HLM or structural equation modeling techniques can be used to produce growth estimates for individual students that account for observations (i.e., testing events) nested within students, and students nested within teachers, schools, and/or districts.

Some true growth VAMs that are generally considered better performing are the univariate value-added model and the multivariate value-added model (Anderman et al., 2015). The univariate value-added response model regresses current student test scores onto previous

test scores in order to create a prediction line for the purpose of comparing actual student performance to expected performance (Anderman et al., 2015). These multilevel models allow for school effects to be included as either fixed or random effects. Additionally, relevant covariates can be included at the student-, teacher-, or school-level. Finally, measurement error has been shown to be mitigated if at least three test scores per student are included (Sanders, 2006). Multivariate value-added response models are more complex and teacher effects from a current year can be layered onto those from prior years. Since these types of models include more information, they have been shown to produce more reliable estimates of student growth and value added by teachers (Hawley et al., 2017; Tekwe et al., 2004).

Anderman and colleagues (2015) summarized the important differences between these types of growth models and their VAM applications. Specifically, they explained that the models that treat school effects as fixed (control for differences across schools) rest on the assumption that schools hire equally capable teachers. Alternatively, models that treat school effects as random not only control for differences across schools, but also hold the assumption that "not all teachers across a district or a region are equally capable," (p. 148) which is a situation more likely to hold true (Anderman et al., 2015).

Regardless of the style of growth model or VAM, research has demonstrated that there can be problems with teacher and school value-added estimates and subsequent classification into categories of growth or school performance ratings. Given that in some states, teacher value-added data can make up nearly 50% of the overall teacher evaluation metrics, which can then be used to make staffing, promotion, or tenure decisions (Blazar et al., 2016; Hawley et al., 2017), issues with estimates and classifications can be quite problematic. Blazar et al. (2016) used a multivariate response model with data from four school districts and found that teachers were

categorized differently when compared within versus between districts. Additionally, the researchers observed large differences in teachers' instructional practices across districts for teachers who received the same within-district value-added ranking. Ng and Koretz (2015) examined the impact of scaling decisions on the sensitivity of school performance ratings (value-added at the school-level). The researchers found that a change in the scaling approach for the assessment resulted in shifts in school rankings and classifications into performance bands. Since most states use different tests (which inherently have different scaling processes) and states typically rescale their tests every few years, these results have implications for the potential decisions that are being made based on value-added results. Importantly, issues have also been raised about the validity of attempting to isolate the true contribution of teachers (i.e., causal effect of teachers on student learning) in a situation without random assignment. Guarino et al. (2015) estimated the teacher effects for a simulated data set that mimicked plausible grouping and teacher assignment scenarios. They found that no single method was able to adequately capture true teacher effects for all scenarios and that the potential for incorrectly classifying teachers was substantial for some models. This implication is quite problematic and highlights why there is not one singularly used VAM across all states.

Since problems arise with the use of growth and VAM results for evaluation of general education teachers, there must also be problems with the use of growth and VAM estimates for the evaluation of teachers whose students have high academic abilities. As previously described, most currently used accountability-driven assessments can be considered minimum competency tests and have been shown to have ceiling effects for students with high academic and intellectual abilities (in addition to various other measurement issues; Callahan, 2009; Lohman & Korb, 2006; Matthews et al., 2013; McBee, 2010; Olszewski-Kubilius, 2010; Plucker &

Callahan, 2014b; Rambo-Hernandez & Warne, 2015; Subotnik et al., 2011; Welsh, 2011). Therefore, it follows that if the assessment cannot capture the true ability of a student, the resulting models of student growth and value-added by teachers must also be problematic (Ryser & Rambo-Hernandez, 2013). Both Ryser and Rambo-Hernandez (2013) and McCoach et al. (2013) emphasized the measurement issues associated with high-ability students taking grade-level assessments that I discussed in the previous section, and they concluded that these measurement issues could cause the predicted growth of these students to be lower than that of average students. Subotnik et al. (2011) also explained that when using standardized instruments (i.e. accountability-driven grade-level standardized achievement tests), high ability student growth estimates were not considered accurate. Welsh (2011) specifically noted that assessments used to measure gains for average students were not useful for capturing the growth of high ability students. Additionally, she explained an important phenomenon, "the amount of gain that can be captured for initially high-performing students will be smaller than for those who started out with lower scores, making their teachers appear less effective" (p. 753).

The inaccuracy of statistical growth modeling when using an assessment with measurement issues is confounded when modeling value-added estimates. Resch and Isenberg (2018) found that with a low test score ceiling, the value-added by teachers of high-ability students shrunk towards the middle of the distribution of value-added estimates for all teachers. Ng and Koretz (2015) also found that the differences in school performance ratings that were attributable to scaling differences were larger when the raw score distribution had ceiling effects. Koedel and Betts (2009) examined the impacts of ceiling effects on value-added estimates as well as teacher classification based on these estimates. The researchers found that imposing

ceiling effects that mimicked minimum-competency testing scenarios significantly impacted value-added rankings.

Ultimately, growth models are being extended and applied as value-added models solely for the purpose of teacher evaluation. In general, it is problematic to use tests to model student growth and teacher value-added contributions when the tests are designed and validated for the purpose of measuring student achievement and classifying students into achievement levels (American Educational Research Association et al., 2014; North Carolina Department of Public Instruction, 2009). However, knowing trends in student achievement is an important piece of information that schools should be using for placement of students into appropriately leveled classrooms, academic acceleration decisions, and differentiation. Therefore, the issues identified here regarding growth models and VAM are primarily associated with their use for teacher evaluation. Modeling the description and prediction of student growth should still be a priority and is the primary focus of this study.

Potential Solutions

The two general problems raised in this review of literature are issues with adequate measurement of high-ability student achievement and adequate modeling of high-ability student growth; therefore, the solutions are directly related. Koedel and Betts (2009) reiterated this issue by explaining the two mechanisms by which ceiling effects distort our understanding of student achievement and growth; specifically, ceiling effects represent lost information about student achievement and result in misspecified models. In order to adequately measure high-ability student academic progress from year-to-year, a more appropriate measure or model is needed. Although the previous section of this literature review dove into the topic of value-added models as they related to accountability issues, the use of student achievement test scores for teacher

evaluation is not a valid interpretation of these tests (given that interpretations of tests are valid and reliable only for their derived purpose and these tests were not written nor studied psychometrically for the purpose of teacher evaluation; American Educational Research Association et al., 2014); therefore, student growth modeling for the purposes of understanding change in student ability from year-to-year is the primary focus of this study.

Above-Level Testing

As previously mentioned, above-level testing has been explored for its uses with high-ability students due to its potential to more accurately capture achievement at the right end of the score distribution. Therefore, a commonly proposed solution in the field of gifted education to problems due to ceiling effects is the use of above-level tests for the measurement of high-ability students' ability (Adelson & Dickenson, 2016; Lee et al., 2008; Subotnik et al., 2011; Swiatek, 2007). Rambo-Hernandez and Warne (2015) detailed the history of above-level testing and provided practical implications for its use. Warne (2012) conducted a comprehensive literature review on above-level testing and found five reasons for the use of such tests, (a) raising the test ceiling, (b) increasing score variability, (c) improving reliability, (d) promoting sound interpretations of above-level test data, and (e) reducing regression to the mean. Above-level testing has been widely used to screen students for talent search programs (Assouline & Lupkowski-Shoplik, 2012; Lee, et al., 2008; Lubinski & Benbow 2006; Swiatek, 2007) and occasionally used to identify students eligible for full-grade acceleration (Assouline et al., 2009; Rogers, 2002). Warne (2014) also used an above-level test to model the growth of high-ability students and found that these students made greater score gains than typical students in the norm group. These results demonstrate the importance of using above-level tests to model high-ability student growth.

Computer-adaptive tests (CAT) are another potential solution that would result in better measurement of high-ability student achievement. CAT are designed to direct students through a unique pathway of items or testlets based on their performance on previous items or testlets in order to gather maximum information about a students' ability level. Adaptive tests are more efficient and precise measurement tools in comparison with fixed-form linear tests where all students are presented with the same set of items along the same difficulty scale (Hendrickson, 2007). As previously explained, fixed-form linear tests have items designed for students of average ability level, therefore these tests are the most accurate measures for only those students. However, these tests are less precise in their measurement of students at the ends of the score distributions (Hambleton & Swaminathan, 1985; Hendrickson, 2007; Lord, 1980; Weiss, 1974). Item-level CAT have been thoroughly researched and shown to allow for shorter tests with at least equivalent measurement precision in comparison with traditional fixed-form tests (Hendrickson, 2007; Lord, 1974; Loyd, 1974; Wainer, et al., 1992).

One specific computer adaptive and above-level assessment that has been used for research purposes in the field of gifted education is the MAP Growth ™ (formerly known as the Measures of Academic Progress; NWEA). This assessment is better able to measure true ability for these students due to its reduction of measurement error typical in grade-level assessments for these students (MAP Growth Technical Report, 2019). The MAP Growth test is administered to the same groups of students up to four time points throughout the year (fall, winter, spring, and summer) and from year-to-year. An in-depth technical explanation of this test will follow in Chapter 3. The MAP Growth test has been frequently used in the field of gifted education for the purposes of measuring and modeling high-ability student growth (McCoach et al., 2021; Rambo-Hernandez et al., 2019; Rambo-Hernandez, Makel, et al., 2021; Rambo-Hernandez & McCoach,

2015; Rambo-Hernandez, Peters, et al., 2021). The technical report also directly explains that the purpose of using a computer-adaptive assessment with items reaching above-grade level when necessary was explicitly for the purpose of gathering more information about students whose ability falls outside of the average range (MAP Growth Technical Report, 2019). Rambo-Hernandez and McCoach (2015) used MAP test results to demonstrate that although high-ability students experienced slower growth during the school year in comparison with their average ability peers, they experienced higher growth during the summer months resulting in an overall steady rate of linear growth. Rambo-Hernandez et al. (2019) used MAP Growth test data to explore the change in excellence gaps over time. The researchers determined through multilevel modeling that excellence gaps were more stable in reading, but increased over time for mathematics. At the American Educational Research Association's 2021 annual meeting, three of the four papers presented in the Research on Giftedness, Creativity, and Talent's paper session "Who Gets to Learn? Student Achievement and Academic Growth" used MAP data to explore issues related to growth based on initial performance, academic diversity within grade-level classrooms, and demographic differences in academic growth based on initial academic proficiency (McCoach et al., 2021; Rambo-Hernandez, Makel, et al., 2021; Rambo-Hernandez, Peters, et al., 2021).

Tobit Modeling

Despite success with above-level and computer adaptive testing, expansive use of such tests for all high-ability students would be expensive for schools and districts to implement. An alternative solution to the use of an above-level test would be the use of an alternate method of modeling student growth. One potential model would be the Tobit model (Tobin, 1958) which was first used to analyze censored outcome variables in econometrics. In general, the Tobit

model refers to the family of regression models whose dependent variables have a constrained range in some way (Amemiya, 1984), and can be estimated using Maximum Likelihood (ML) or Bayesian techniques (Cowles et al., 1996). The constraint on the dependent variable can be through censoring or truncation. When the dependent variable is censored (either from above, below, or both), the dependent variable is observed for all values of the independent variable, but the outcome is suppressed to the censor point(s). The dependent variable is considered truncated when there are no observations of the dependent variable past the truncation point in the independent variable. Anecdotally, an example of a scenario with a censored dependent variable could be when students hit the measurement ceiling of an assessment and therefore can only receive the highest possible score, despite their potential for scoring higher if the assessment included above-grade level material. An example of a truncated dependent variable situation could be if test scores for students whose scores fall above a certain threshold were not included in the data set at all. In the present study, the dependent variable is censored rather than truncated.

There are two main types of Tobit models. The Type I Tobit model can be considered a combination of a probit model (using the probability of being censored) for the censored observations and a truncated regression model for when the dependent variable is not censored. This type of Tobit model is estimated most frequently using ML estimation methods. The Type II Tobit model occurs when the independent variable has different effects on the uncensored and censored portions of the data. The estimation of this type of Tobit model is best completed with a Heckman two-step estimator (Heckman, 1974; 1976). This study uses exclusively Type I Tobit models.

Tobit Estimation and Correction

ML Estimation of the Tobit model requires two steps, first the observations are split into two groups, censored and uncensored. The uncensored observations are treated in the traditional way and estimates are produced. Since the true values of the censored observations are unknown, the probability of being censored is used in the likelihood equation to create a weighted regression that is used to predict uncensored scores for the censored cases, under the assumption of normality of the dependent variable. Cases with a higher probability of being censored receive higher corrected estimates (See Appendix A for the correction procedure). Equation 3 shows the basic single level Tobit model with censoring from above.

$$y_i = \begin{cases} \boldsymbol{x_i\beta} + \varepsilon_i & if\ y*_i < \tau \\ \tau_y & if\ y*_i \geq \tau \end{cases} \quad (3)$$

Where errors ε_i are normally distributed, $\boldsymbol{x_i}$ is a vector of independent variables, $\boldsymbol{\beta}$ is a vector of regression coefficients, $y*_i$ is a latent variable that is observed for all values less than τ and is censored for all values greater than or equal to τ, τ is the threshold that determines whether or not $y*_i$ is censored, and τ_y is the value assigned to y_i when $y*_i$ is censored (Breen, 1996; Long, 1997). The Tobit model has also been expanded to allow for a hierarchical structure.

Methodological Research on Tobit Modeling

The Tobit model has been explored for its methodological soundness and advantages over other modeling strategies by a few authors in the field of statistics and methodology. These methodological articles primarily center around implications of violations of assumptions, performance of the model under certain conditions and in comparison to other estimation procedures, and extensions beyond its initial uses.

There are two primary assumptions that must hold true for the Tobit model's corrections to be accurate; the errors must be normally distributed and homoscedastic, and the dependent variable must be normally distributed (albeit incompletely observed at the censor point).

Multiple studies have explored the robustness of the Tobit model under conditions that do not meet these assumptions and have found the estimates produced by the model to be biased and inconsistent (Abramazar & Schmidt, 1981; Abramazar & Schmidt, 1982; Caudill & Mixon, 2009; Holden, 2004). Therefore, it is extremely important that data meet these requirements if Tobit modeling is to be used.

Additionally, methodological research has explored the performance of the Tobit model compared with other estimation procedures. Baba (1990) compared the Tobit model to traditional ordinary least squares (OLS) when using a truncated distribution and found that the Tobit model resulted in additional explained variance over the basic OLS procedure. Brown and Dunn (2011) compared Tobit, linear, and Poisson-gamma regression models with simulated time use data and found the Poisson-gamma distribution to be more empirically sound based on factors such as the interpretation of the model and model fit (residual analysis); however, the Tobit regression model did perform better than the basic linear regression models. Finally, Leiker (2009) compared estimates produced by the Tobit procedure with those produced by Probit ML estimators, Least Squares estimators, the Heckman Two-Step estimator, and the Expectation Maximization algorithm. The researcher found that in eleven of the twelve simulations conducted, the Tobit ML estimation procedure produced the most accurate estimates and had the smallest errors.

Research on the Tobit model has also explored extensions beyond its primary uses. Solon (2010) developed a microeconomic theory for specifying Tobit models under certain conditions of consumer demand. Wright & Ziegler (2015) explored the potential for extending single censor point Tobit models to include data that are censored at multiple points. The authors also created a package in R to conduct these analyses. Despite its promise, the article has not

been widely cited, nor the package widely used. Wang and Zhang (2011) used Tobit modeling to explore the influences of censored data on mediation effects using structural equation modeling. Finally, Wang and Griswold (2017) explored the modeling of overall exposure effects using a direct-marginalization approach with censored or truncated dependent variable means through simulation studies.

Applied Research on Tobit Modeling

Applied research using Tobit modeling has primarily come from the field of econometrics, which is unsurprising given its origin in that field. McDonald and Moffit (1980) provided an overview of Tobit modeling for an economic-focused audience and explored the disaggregation of Tobit effects. Shishko and Rostker (1976) studied the economics of multiple job holding, a dataset appropriate for Tobit modeling since there was a large cluster of observations at zero, due to many people not holding a second job. Grootendorst (1997) conducted an evaluation of health care policy, specifically pharmaceutical costs, using panel data; these data were censored when participants began paying reduced costs for pharmaceuticals as they aged into the Medicare system. Barros et al. (2018) provided a methodological overview of Tobit and discussed potential applications for the field of econometrics. Amore & Martinu (2019) also provided an overview of Tobit modeling as well as the frequency of the model's use in the field of econometrics. Additionally, they conducted an example application of Tobit modeling by demonstrating the effects of foreign competition on diversification of corporate portfolios.

The field of health and human sciences has also produced several applied studies. Austin et al. (2000) provided an overview of potential uses for Tobit modeling in the field of health science by demonstrating its use on simulated health status data with ceiling effects (or right

censored). Delva et al., (2006) used Tobit modeling to explore youth alcohol problems based on youth depression and parental factors. Twisk & Rijmen (2009) conducted a longitudinal (multilevel) Tobit regression on epidemiological data with both floor and ceiling effects and found that it performed better than the traditional mixed effects model.

Additionally, fields such as political science, veterinary medicine, and agricultural science have also produced a few applied studies that used Tobit modeling. Sigelman and Zeng (2000) discussed the methodology of Tobit modeling and provided practical situations where its use would be necessary in the field of political science. Ekstrand and Carpenter (1998) extended the use of Tobit modeling into the field of veterinary medicine by studying the risk factors for foot-pad dermatitis in poultry livestock. Allcroft and Glasby (2003) introduced Tobit modeling into the field of agricultural research in their study of crop lodging with a high number of observations at zero (left censored).

Ultimately, Tobit modeling has yet to be fully applied in the field of education. Koedel and Betts (2009) briefly examined a Tobit model as a secondary focus of their full study and found that Tobit specification improved model performance substantially, with emphasis on its performance in a minimum-competency simulation. Although the focus of this study was in the field of education, this exploration could still be considered a methodological study rather than applied since it used simulated data. McBee (2010) also simulated data to demonstrate the effectiveness of Tobit modeling at producing estimates from artificially censored data that matched his initial simulated (uncensored) data. Most similar to this study, though also using simulated data, Wang et al. (2008) explored the effectiveness of Tobit modeling in a longitudinal growth modeling scenario (using a structural equation modeling framework) and found that Tobit modeling performed well when dealing with ceiling effects. Despite its potential for use

when modeling the growth of high-ability students, Tobit modeling has not been applied to such problems.

Research Questions

This study sought to explore the options available for solving the problems associated with the use of accountability-driven grade-level standardized achievement assessments with high-ability students. In order to demonstrate the true growth of high-ability students using an above-level assessment and use Tobit modeling with a grade-level assessment to demonstrate true growth for high-ability students, two main research questions must be answered.

1. Can multilevel Tobit regression be used to replicate true growth estimates on artificially censored data?
    a. At what level of achievement is multilevel Tobit regression preferable over traditional multilevel modeling procedures?
2. Are growth estimates from Tobit Regression significantly different from those obtained using traditional models on naturally censored data?

CHAPTER 3: METHOD

This study comprises two parts, hereafter referred to as Part One and Part Two. Part One used MAP Growth test data and Part Two used North Carolina End-Of-Grade (NC EOG) test data. This chapter provides detailed psychometric information about each of the two assessments and discusses the method of data analysis. The chapter concludes with a brief discussion related to ethical considerations.

Measures

MAP Growth

MAP Growth assessments are untimed, interim, computer-adaptive tests administered to students in grades K-12 (K-2 is a separate test) with the purpose of measuring student achievement and growth in Reading and Mathematics (in addition to Language Use and Science). Typically, the tests are administered at three time points throughout the year (fall, winter, spring) with an optional summer administration. The use of this assessment has several advantages over typical state-administered tests given at the end of each grade level. The MAP Growth test is computer-adaptive, which means that as students progress through the test, they are presented with appropriately challenging items (which can span grade-levels depending on student performance) designed to gather the most information about a student's ability level.

MAP Growth scores are reported using a Rasch Unit (RIT) scale. RIT scores are continuous and on an equal-interval scale which make tracking progress over time easily interpretable and meaningful. RIT scores were developed using a "one-parameter Rasch IRT model that estimates the probability that a student with an achievement score of θ will correctly answer a test item of difficulty δ" (MAP Growth Technical Report, 2019, p. 53). The achievement score and item difficulty are both expressed on the logit metric and the following

transformation is performed to create the RIT scale (See Equation 4). Scores range from 100 to 350.

$$RIT = (\theta * 10) + 200 \quad (4)$$

The use of this Rasch model-derived scale has multiple benefits. Specifically, item difficulty and ability estimation are not sample-dependent and differ only in measurement precision. Additionally, "item difficulty values define test characteristics. This means that once the difficulty estimates for the items to be used in a test are known, the precision and the measurement range of the test are determined" (p. 54).

MAP Growth tests draw from an item bank containing over 42,000 expert-written items designed to align with Common Core State Standards or state-specific standards when necessary. All items are multiple-choice or technology-enhanced items and are dichotomously scored. New items are continuously in development through an eight-step item writing, review, and field-testing process. The steps are as follows: determine item needs, write specifications, write items, validate item content, item content review, second content review, item quality review, and field testing. The item bank is also consistently reviewed for quality of alignment, content accuracy, relevance, bias, and sensitivity. Additionally, alignment studies are frequently conducted by both NWEA as well as third parties to ensure proper alignment of items to standards. A recent alignment study found all items to have "very good alignment in terms of categorical concurrence, cognitive complexity, and range and balance of knowledge" (MAP Growth Technical Report, p. 18; Egan & Davidson, 2017).

Tests are constructed by limiting the item pool based on specific criteria including instructional areas and sub-areas, standards aligned to these areas, item selection requirements, and item filters based on specific item metadata. Students are assigned a starting value based on

previous test performance or (if no prior test scores exist) based on grade level. Item selection occurs throughout the test based on interim Bayesian-estimated student ability level, content requirements, and item exposure controls. As students progress through the test, their ability estimate becomes more precise and the test ends when the standard error reaches its minimum obtainable value. Tests go through multiple rounds of checks to be sure that each combination of items meets validity and reliability requirements. Specific attention is paid to depth of content by simulating test-student run-throughs for various ability levels to ensure adequate achievement continuums are covered. Tests also endure a rigorous content validation process including the analysis of multiple simulation studies conducted by NWEA psychometricians.

Reliability of MAP Growth tests was examined using test-retest reliability, marginal reliability, and score precision. Test-retest reliability answers the question, “To what extent does the test administered to the same students twice yield the same results from one administration to the next?” (p. 82). Due to the adaptive nature of the MAP Growth test, test-retest reliability presented in the technical report is a mix of test-retest reliability and a type of alternate forms reliability. Practically, it describes the influence of time and item selection which are the two main sources of measurement error. Marginal reliability answers the question, “to what extent do items in the test measure the test’s construct(s) in a similar manner?” (p. 82). This type of reliability is an equally valid alternative to internal consistency, which the technical report describes as cumbersome and inaccurate for adaptive tests. Finally score precision is examined via the SEM of scores. Lower SEM indicates higher test information and higher precision; with an adaptive test, SEMs should be comparable across a wide achievement range. The 2019 technical report included data from Fall 2016, Winter 2017, Spring 2017, and Fall 2017 from testing events from all 50 states plus the District of Columbia. The states included in this analysis

were anonymized, however all reliability statistics were within the acceptable range (> .70). Test-retest alternate forms reliability evidence indicates that MAP Growth scores are highly consistent for students across both states and grade levels. Marginal reliability statistics were also consistent across states and demonstrated that measurement error was only a minor portion of the overall score variability. Finally, the distribution of SEM of test scores was consistent across the RIT scale except at the very high and very low levels, which was expected. Although these SEMs are slightly higher for high-ability students, they are far lower than what would be expected from fixed-form grade level assessments.

Validity of MAP Growth tests is an ongoing process throughout the entire testing process all the way through the interpretation and use of scores. During the item-writing steps, an internal review process for content standards and item quality requires content specialists to collect data to examine content validity. Additionally, third parties conduct alignment studies to address the content validity as it relates to CCSS. Criterion-related validity includes both concurrent validity as well as classification accuracy statistics. Concurrent validity is expressed as a Pearson correlation between MAP Growth RIT scores and other established and validated state test scores. Classification accuracy statistics are based on predicted proficiency on state summative assessments. To predict proficiency, linking studies are conducted to establish MAP Growth cut scores that correspond to proficiency on state tests. Accurate classification occurred for 83% of reading students and 87% of mathematics students, indicating that MAP Growth cut scores established by linking studies are effective predictors of student proficiency on state assessments.

Additionally, the technical report describes validity evidence for the internal structure of the test in its alignment with theoretical expectations, test design, and differential item

functioning (DIF). NWEA has conducted studies that examined underlying constructs in the tests and found that these constructs remained consistent across grade levels (Wang, Jiao, et al., 2013; Wang, McCall, et al., 2013). This evidence supports the use of test results for measuring student achievement across time. Another type of validity evidence was provided by results from test-taking engagement studies. Since this is a computer adaptive test, students are given appropriately challenging questions that tend to keep them more engaged, thus reducing measurement error and validating the use of these results. Finally, differential item functioning procedures (Mantel-Haenszel, 1959; Walker, 2011; Zwick et al., 1999) examined item performance across reference and focal groups to ensure validity of results for different ethnicities and genders.

North Carolina End-Of-Grade Test

North Carolina's EOG tests are timed, summative, fixed-form paper-and-pencil or computer-administered multiple-choice assessments. Students in grades 3 through 8 take EOG tests for Mathematics and English Language Arts, and students in grades 5 and 8 take an additional EOG in Science. Additionally, during the high school years, certain courses require summative End-of-Course exams. The statewide testing program was initiated for tracking student achievement for the purpose of accountability.

Scores are reported as scale scores, percentiles, and achievement levels. Scale scores are derived from summed raw scores and are expressed along a vertical scale. Percentile ranks compare students to the performance of other students who took the test during the norming year (Spring 2008). For the years of data used in this study (2008-2011), there were four achievement levels, though in 2012 the tests were rescaled and adapted to have five achievement levels. Achievement level I was defined as insufficient proficiency, level II indicated inconsistent

proficiency, level III was deemed consistent grade-level proficiency, and level IV indicated consistent superior performance beyond grade level. As previously discussed, these tests do not include any above grade-level content, so this description of proficiency level IV is slightly deceptive. Cut scores for different achievement levels were created using the bookmark method and these cut scores were used for the first time in the Spring 2008 administration and were used continuously until the scores were re-normed and the scale adjusted in 2012.

The test development process followed a 22-step sequence of events which took place over the course of 49 months. Steps involved item development with multiple stages of item reviews, field testing with multiple stages of review, parallel form construction and review, pilot testing with more review stages, test scoring and standard setting, and finally the administration of the fully operational test and associated reporting of results. Item writers were recruited from a pool of educators with strong knowledge of the NC Standard Course of Study (NCSCS) as well as their content area and diversity. Educators were used in order to establish instructional validity. Items were written to align with the NCSCS since the state had not adopted Common Core State Standards prior to 2011. Item writers were guided by a test blueprint with specifications for item types, thinking level, difficulty level, and specific content objectives. After several rounds of reviewing and field testing, the item pool was finalized. Items statistics were reviewed based on CTT (point-biserial correlation and p-value), IRT (three-parameter logistic model), and DIF (Mantel-Haenszel procedure) to ensure only high-quality items were retained in the final versions of each test form. Items were retained in the pool if they met the following criteria: p-value near 0.625, a parameter (slope) greater than .60, c parameter (asymptote) less than .40, and DIF log odds ratio between 0.67 and 1.5.

NC EOG test reliability evidence was examined using coefficient alpha as a measure of internal consistency. All coefficients were greater than .82 which is generally accepted as good reliability. SEMs were also examined with SEMs greater at score extremes (low and high) for each grade level. To ensure that parallel forms were truly parallel, test characteristic curves were graphed on top of each other for each grade level and subject. Since each curve falls directly onto or is very closely aligned with others, differing test forms appear to be essentially parallel.

Validity evidence was presented for content validity, instructional validity, and criterion-related validity. To establish content validity, the technical report displayed tables showing content goals from NCSCS and the number of corresponding questions per goal. Additionally, they highlighted the use of educators as item-writers and their familiarity with the content standards as evidence of content validity. Instructional validity was established through teacher surveys during the test development process. Teachers were asked to comment on items that did not reflect goals and objectives of the curriculum or items whose content did not appear equitable across diverse groups. Criterion-related validity evidence presented in the technical report included correlations between student raw scores and teacher predicted course grades and achievement levels. The correlations ranged from .50 to 0.69 which indicated moderate correlations.

Participants and Samples

MAP Growth

Data were provided by NWEA via the Kingsbury Research Award. Specific data selected for use in this study were based on several criteria. Data from the five states with the highest MAP Growth participation and North Carolina (all of which were anonymized in the data set) were included. Cohorts of students who were in third or sixth grade in the 2016-2017 school year

were chosen to have their growth tracked for three years (ending in Spring 2019) so that data were not impacted by the COVID-19 pandemic. Table 1 shows descriptive statistics conveying the demographic makeup of the whole sample.

Table 1

MAP Growth Demographics

	Elementary Mathematics		Elementary Reading		Middle School Mathematics		Middle School Reading	
	n	%	n	%	n	%	n	%
Gender								
Male	165653	50.9	166350	50.9	153598	50.9	152903	50.9
Not Male	159642	49.1	160298	49.1	148079	49.1	147568	49.1
Title I Status								
Title I	212719	65.4	214033	65.5	178755	59.3	177894	59.2
Not Title I	112576	34.6	112615	34.5	122922	40.7	122577	40.8
Minority Status								
Minority	166224	51.1	167037	51.1	148343	49.2	147624	49.1
Not Minority	159071	48.9	159611	48.9	153334	50.8	152847	50.9
Ethnicity								
American Indian or Alaska Native	2319	0.7	2324	0.7	1994	0.7	1992	0.7
Asian	11669	3.6	11565	3.5	10605	3.5	10567	3.5
Black or African American	72029	22.1	73099	22.4	61313	20.3	61291	20.4
Hispanic or Latino	47270	14.5	47035	14.4	41859	13.9	41777	13.9
Native Hawaiian or Other Pacific Islander	369	0.1	366	0.1	330	0.1	331	0.1
White	159071	48.9	159603	48.9	153328	50.8	152847	50.9
Multi-Ethnic	11463	3.5	11622	3.6	9484	3.1	9415	3.1
Not Specified or Other	21096	6.5	21026	6.4	22758	7.5	22235	7.4
Total	325295	100	326648	100	301677	100	300471	100

The most recent norming report was published by NWEA in 2015. These norms were created for grades K-12 using testing events from fall, winter, and spring from 2011 to 2014 from all 50 states plus the District of Columbia. Post-stratification population weights were constructed using a school challenge index calculated based on several National Center for Education Statistics demographic variables such as school locale, school type, title I status, proportion free or reduced lunch, proportion of ethnic minority, and full-time equivalent classroom teachers.

North Carolina End-Of-Grade Test

North Carolina EOG Test data were provided by the NCERDC. Cohorts of students who were in third or sixth grade during the 2008-2009 school year were selected to avoid the rescaling that occurred with both the mathematics and reading tests in 2012. Table 2 shows the demographic makeup of the whole sample.

Table 2

NC EOG Demographics

	Elementary Mathematics		Elementary Reading		Middle School Mathematics		Middle School Reading	
	n	%	n	%	n	%	n	%
AG Status								
AG	4786	3.6	4703	3.5	11412	9.2	10784	8.7
Not AG	128707	96.4	12890	96.5	112500	90.8	113128	91.3
Gender								
Male	59887	44.9	59880	44.9	60104	48.5	60101	48.5
Not Male	73606	55.1	73613	55.1	63808	51.5	63811	51.5
Title I Status								
Title I	69910	52.4	69911	52.4	19041	15.4	19041	15.4
Not Title I	63583	47.6	63582	47.6	104871	84.6	104871	84.6
Minority Status								
Minority	55313	41.4	55309	41.4	53968	43.6	53976	43.6
Not Minority	78180	58.6	78184	58.6	69944	56.4	69936	56.4
Ethnicity								
Asian	2935	2.2	2935	2.2	3071	2.5	3069	2.5
Black	31155	23.3	31147	23.3	32655	26.4	32659	26.4
Hispanic	14042	10.5	14045	10.5	12022	9.7	12025	9.7
American Indian	1826	1.4	1826	1.4	1718	1.4	1719	1.4
Multi Racial	5355	4.0	5356	4.0	4502	3.6	4504	3.6
White	62274	46.6	62278	46.7	63431	51.2	63427	51.2
Not Specified or Other	15906	11.9	15906	11.9	6513	5.3	6509	5.3
Total	133493	100	133493	100	123912	100	123912	100

Data Analysis

MAP Growth

Part one of this study used MAP Growth data to explore the use of Multilevel Tobit Regression with artificially censored data. To accomplish this, I first created sub-samples from each cohort consisting of students below the 95th percentile, students at or above the 95th percentile, and students at or above the 98th percentile. With each sample, cohort, and subject, I ran multilevel mixed-effects growth models with measurements nested within students across three years (up to nine time points) to establish "true" monthly growth rates. Next, I created an artificial ceiling by censoring scores above the thresholds determined using the results of the linking study done by NWEA (NWEA Psychometric Solutions, 2021) that linked MAP Growth cut scores with proficiency levels and scale scores on NC EOGs. Third, I ran a multilevel Tobit regression using the *metobit* function in Stata with these artificially censored test scores as the dependent variable to attempt to replicate the original "true" growth estimates. All of the models included the covariates gender (coded 1 for male), a dichotomous variable that indicated whether or not a student attended a school that qualified as Title I, and a dichotomous variable that indicated whether or not a student was a minority. Additionally, interactions between the demographic variables and time were included. Those covariates were chosen based on models used in previous studies using the same data sets and with similar goals (Rambo-Hernandez & McCoach, 2015; Rambo-Hernandez et al., 2021). The results of these models will be detailed in the next chapter. The basic form of the multilevel equations is shown in Equation 5.

$$\begin{gathered} ASSESS\_RIT_{ti} = \pi_{0i} + \pi_{1i} * TEST\_TIME_{ti} + e_{ti} \\ \pi_{0i} = \beta_{00+}\beta_{01} * MALE_i + \beta_{02} * TITLE_{I_i} + \beta_{03} * MINORITY_i + \beta_{04} * (TIMExMALE)_i + \beta_{05} \\ * (TIMExTITLEI)_i + \beta_{06} * (TIMExMINORITY)_i + r_{0i} \\ \pi_{1i} = \beta_{10} + r_{1i} \end{gathered} \tag{5}$$

North Carolina End-Of-Grade Test

Part two of this study used NC EOG test data to attempt to apply the usage of Tobit regression to naturally censored data. For these data, sub-samples were created by separating

students who were and were not identified as Academically Gifted (AG) in either reading or math into their own files. First, I ran a multilevel growth model with measurements nested within students across three years for mathematics and reading for both the elementary and middle school cohorts (using the samples including all students, the subsample containing only AG students, and the subsample containing only non-AG students). These data are generally accepted as naturally censored for students scoring two standard deviations above the mean due to the lack of above grade-level content (Plucker & Callahan, 2014b; Olszewski-Kubilius, 2010; Subotnik et al., 2011; Welsh, 2011). Second, I ran a multilevel Tobit regression using *metobit* in Stata with these same samples and compared the resulting growth estimates with the original estimates. Recall from chapter 2 that CSEM is higher at the ends of the ability distribution, and therefore it is likely that the scores received by the highest achieving students (those scoring at or above the 95th percentile) could have been at the ceiling on a different testing occasion due to random measurement error. To account for this, I lowered the ceiling to the 95th percentile and ran the multilevel Tobit regression again. Similar to the first part of the study, the covariates gender (male), Title I status, and minority were included as well as each covariates' interaction with time. These results will be discussed in the next chapter. The general form of the multilevel equations is shown in Equation 6.

$$SCALE\_SCORE_{ti} = \pi_{0i} + \pi_{1i} * TEST\_TIME_{ti} + e_{ti}$$
$$\pi_{0i} = \beta_{00+}\beta_{01} * MALE_i + \beta_{02} * TITLE_{I_i} + \beta_{03} * MINORITY_i + \beta_{04} * (TIMExMALE)_i + \beta_{05} * (TIMExTITLEI)_i + \beta_{06} * (TIMExMINORITY)_i + r_{0i}$$
$$\pi_{1i} = \beta_{10} + r_{1i} \quad (6)$$

Ethical Considerations

Prior to beginning analyses, IRB approval was obtained for both data sets. I obtained access to the NC EOG data from the North Carolina Education Research Data Center. I was also awarded access to the MAP Growth data via the Kingsbury Research Award from NWEA. Both

NC ERDC and NWEA have requested to see final manuscripts prior to publication, however they have stipulated that only incorrect usage of their data would result in required edits to the manuscript.

CHAPTER 4: RESULTS

The results that follow are organized into two parts to reflect the design of the study. The first part details the results of the models run using MAP Growth data. The second part details the results of the models run using NC EOG test data. Tables 5, 7, 9, and 11 show the in-depth model results, while the narrative contains highlights and explanations. Maximum likelihood estimation requires that the errors are both homoscedastic and normal for the Tobit model to produce unbiased estimates. These assumptions were visually checked by plotting residual plots for each sample, cohort, and subject of data. The plots demonstrated that the residuals were random, and therefore the assumption was met. Additionally, the Tobit model requires normal distribution of the data, barring the portion that has been censored; all samples, cohorts, and subjects of data were visually inspected and all had normal distributions.

Part One: MAP Growth

Three models were run for each subject, cohort, and sample of data. Model 1 used the uncensored test scores as the outcome and time after testing (in months) as the predictor (in addition to the demographic variables gender, Title I status, and minority status), with clustering occurring at the student level. Model 2 used the same predictors and clustering structure, but used the artificially censored test scores as the outcome. Model 3 used the same predictors and clustering structure as the previous two models, but used multilevel Tobit regression rather than traditional ML multilevel modeling; the censor point for the model was set at the upper limit, despite censoring occurring at up to nine time points across the data. For the sample of students scoring at lower than the 95th percentile, only models 1 and 2 were run; since there was no censoring in those samples, Tobit regression was not necessary. Table 3 shows the percentages

of censored observations at the artificially imposed test ceiling disaggregated by sample, subject, and cohort.

Table 3

MAP Growth Percent of Observations Censored at the Ceiling

Sample and Cohort	%
Whole Sample	
Elementary Mathematics	0.43
Elementary Reading	0.23
Middle School Mathematics	0.88
Middle School Reading	0.14
Less than 95th Percentile	
Elementary Mathematics	< 0.01
Elementary Reading	0.02
Middle School Mathematics	0.19
Middle School Reading	< 0.01
95th Percentile	
Elementary Mathematics	9.48
Elementary Reading	7.28
Middle School Mathematics	13.98
Middle School Reading	7.18
98th Percentile	
Elementary Mathematics	15.81
Elementary Reading	6.35
Middle School Mathematics	14.68
Middle School Reading	9.25

Several overarching trends emerged from the results of this portion of the study. First, the use of Tobit regression did make differences in the growth estimates for particular samples of students, despite not perfectly replicating uncensored growth estimates. Second, the models using the samples of students scoring at the 95th percentile or higher were best suited for Tobit regression, evidenced by the models' ability to come close to replicating the estimates produced using the uncensored test scores as the outcome. At lower levels of achievement there were no censored cases, making Tobit regression unnecessary. At the 98th percentile of achievement, the observations were more heavily censored and therefore Tobit correction procedures did not come as close to the uncensored estimates as it did for the 95th percentile sample. Finally, although not an explicit research question, it is worth noting that students performing at higher levels of achievement demonstrated higher levels of growth than their lower performing peers.

Whole Sample

The whole sample had the following means and standard deviations across all time points, 205.23 (*SD* = 27.5), 201.00 (*SD* = 17.56), 224.91 (*SD* = 18.71), and 217.15 (*SD* = 16.08) for elementary math and reading and middle school math and reading, respectively. In both cohorts the average achievement was slightly higher for math than for reading. The average achievement at each time point for both the uncensored and censored outcome variables for the entire sample is displayed in Table 4.

Table 4

MAP Growth Mean Achievement Disaggregated by Time- Whole Sample

	Elementary Math				Elementary Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall ‘16	189.04	13.50	188.92	13.21	187.83	17.09	187.77	16.96
Winter ‘16	195.54	13.17	195.46	12.95	193.84	16.42	193.80	16.33
Spring ‘17	202.19	14.19	201.98	13.73	198.63	16.41	198.55	16.25
Fall ‘17	200.61	14.13	200.47	13.80	197.55	16.54	197.49	16.43
Winter ‘17	205.98	14.05	205.84	13.72	202.46	15.61	202.43	15.53
Spring ‘18	212.13	15.65	211.83	15.02	205.81	15.62	205.74	15.48
Fall ‘18	210.32	15.48	210.17	15.15	205.23	15.95	205.16	15.80
Winter ‘18	214.96	15.78	214.81	15.42	209.08	15.13	209.02	15.01
Spring ‘19	220.49	17.43	220.18	16.78	211.52	15.26	211.42	15.07
	Middle School Math				Middle School Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall ‘16	215.17	16.01	215.03	15.71	210.64	16.23	210.62	16.20
Winter ‘16	218.83	16.40	218.71	16.14	213.36	15.63	213.34	15.59
Spring ‘17	223.30	17.49	223.12	17.12	215.86	15.85	215.82	15.78
Fall ‘17	222.02	17.50	221.85	17.12	215.02	15.97	214.98	15.89
Winter ‘17	225.30	17.72	225.13	17.35	217.61	15.35	217.56	15.24
Spring ‘18	229.30	18.60	229.03	18.04	219.79	15.38	219.72	15.22
Fall ‘18	228.26	18.42	227.85	17.62	219.53	15.54	219.49	15.45
Winter ‘18	231.00	18.73	230.55	17.86	221.64	15.16	221.59	15.05
Spring ‘19	234.27	19.64	233.58	18.39	223.13	15.23	223.08	15.10

The results displayed in Table 5 are the results of models using the whole sample for each subject and cohort. Growth estimates were mostly unchanged when comparing the artificially censored models to the multilevel Tobit models when using the whole group sample. When examining the changes in the multilevel models, it was important to take into consideration the

percent of observations censored only at the measurement ceiling. The MAP Growth data sets had approximately .14% to .88% of observations censored at the ceiling. Given the low percentages of censored observations, it was unsurprising that estimates did not change much between models. It is important to remember, however, that the MAP Growth estimates represent monthly growth rates; therefore, very small changes in MAP Growth estimates would amount to slightly larger changes across the entire year (approximately ten times larger, given a ten-month school year). Intraclass correlation coefficients (ICCs) ranged from .82 to .89, indicating a large amount of variance was accounted for by the hierarchical structure.

Table 5a

MAP Growth Model Results- Elementary Whole Sample

	Elementary Mathematics			Elementary Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	193.51 (.04)***	193.46 (.04)***	193.41 (.04)***	195.52 (.05)***	195.46 (.05)***	195.40 (.05)***
Time	0.89 (.001)***	0.89 (.001)***	0.89 (.001)***	0.63 (.001)***	0.63 (.001)***	0.63 (.001)***
Male	0.75 (.05)***	0.63 (.05)***	0.62 (.05)***	-2.73 (.05)***	-2.73 (.05)***	-2.72 (.05)***
Title I	-1.57 (.02)***	-1.62 (.02)***	-1.57 (.02)***	-2.75 (.02)***	-2.77 (.02)***	-2.72 (.02)***
Minority	-5.84 (.05)***	-5.81 (.05)***	-5.79 (.05)***	-7.65 (.05)***	-7.62 (.05)***	-7.59 (.05)***
Time X Male	0.02 (.001)***	0.01 (.002)***	0.02 (.001)***	0.004 (.001)***	0.004 (.001)***	0.004 (.001)***
Time X Title I	0.51 (.002)***	0.51 (.002)***	0.50 (.001)***	0.53 (.002)***	0.53 (.002)***	0.53 (.002)***
Time X Minority	-0.06 (.001)***	-0.06 (.001)***	-0.06 (.001)***	0.02 (.001)***	0.02 (.001)***	0.02 (.001)***
Variance Components						
Intercept	179.49	169.09	169.67	205.52	202.02	202.62
Residual	35.92	34.27	34.52	45.73	45.2	45.31
ICC	0.83	0.83	0.83	0.82	0.82	0.82

Note: * p < .05, ** p < .01, *** p < .001

Table 5b

MAP Growth Model Results- Middle School Whole Sample

	Middle School Mathematics			Middle School Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	219.01 (.05)***	219.16 (.05)***	219.03 (.05)***	215.95 (.05)***	215.92 (.05)***	215.89 (.05)***
Time	0.58 (.001)***	0.56 (.001)***	0.57 (.001)***	0.35 (.001)***	0.35 (.001)***	0.35 (.001)***
Male	0.42 (.06)***	0.35 (.06)***	0.34 (.06)***	-2.87 (.05)***	-2.86 (.05)***	-2.86 (.05)***
Title I	-0.31 (.02)***	-0.58 (.02)***	-0.44 (.02)***	-0.54 (.02)***	-0.54 (.05)***	-0.51 (.02)***
Minority	-7.43 (.06)***	-7.48 (.06)***	-7.44 (.06)***	-6.72 (.05)***	-6.71 (.05)***	-6.70 (.05)***
Time X Male	-0.04 (.001)***	-0.04 (.001)***	-0.04 (.001)***	-0.04 (.001)***	-0.04 (.001)***	-0.04 (.001)***
Time X Title I	0.27 (.002)***	0.29 (.002)***	0.28 (.002)***	0.20 (.002)***	0.20 (.002)***	0.20 (>002)***
Time X Minority	-0.03 (.001)***	-0.03 (.001)***	-0.03 (.001)***	0.04 (.001)***	0.04 (.001)***	0.04 (.001)***
Variance Components						
Intercept	275.04	256.78	258.64	198.25	195.87	196.19
Residual	33.50	32.00	32.36	41.15	40.86	40.94
ICC	0.89	0.89	0.89	0.83	0.83	0.83

Note: * p < .05, ** p < .01, *** p < .001

Lower Than 95th Percentile Achievement

These subsamples of students had the following average achievement across all time points, 203.77 (*SD* = 16.21), 200.16 (*SD* = 17.05), 223.05 (*SD* = 17.14), and 216.41 (*SD* = 15.53) for elementary math (n = 321,356) and reading (n = 324,892), and middle school math (n = 296,193) and reading (n = 297,775), respectively. Table 6 shows descriptive statistics for this sample at each time point for each subject and cohort. The estimates produced by models 1 and 2 using this sample of students did not differ and therefore Tobit regression (model 3) was not necessary (See Table 7 for the full model results). The ICCs for this sample of students were very similar to that of the whole group sample, simply because this sub-sample made up most of the students in the whole sample.

Table 6

MAP Growth Mean Achievement Disaggregated by Time- Less than 95th Percentile Sample

	Elementary Mathematics				Elementary Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall '16	187.83	12.46	187.83	12.46	186.33	16.03	186.33	16.03
Winter '16	194.71	12.37	194.71	12.37	192.89	15.73	192.89	15.73
Spring '17	200.87	13.03	200.87	13.03	197.55	15.65	197.55	15.65
Fall '17	199.41	13.05	199.41	13.05	196.53	15.80	196.53	15.80
Winter '17	204.81	12.88	204.81	12.88	201.78	15.09	201.78	15.09
Spring '18	210.32	14.05	210.32	14.05	205.05	15.05	205.05	15.05
Fall '18	208.63	14.03	208.63	14.03	204.41	15.34	204.41	15.34
Winter '18	213.37	14.30	213.37	14.30	208.52	14.69	208.52	14.69
Spring '19	218.23	15.49	218.23	15.49	210.89	14.76	210.89	14.76
	Middle School Mathematics				Middle School Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall '16	213.76	14.80	213.76	14.80	209.88	15.67	209.88	15.67
Winter '16	217.60	15.33	217.60	15.33	212.74	15.16	212.74	15.16
Spring '17	221.94	16.38	221.94	16.38	215.13	15.32	215.13	15.32
Fall '17	220.16	16.00	220.16	16.00	214.22	15.39	214.22	15.39
Winter '17	223.58	16.31	223.58	16.31	216.89	14.80	216.89	14.80
Spring '18	227.26	16.95	227.26	16.95	219.00	14.77	219.00	14.77
Fall '18	226.11	16.67	226.11	16.67	218.75	14.92	218.75	14.92
Winter '18	228.83	16.92	228.83	16.92	220.92	14.57	220.92	14.57
Spring '19	231.60	17.47	231.58	17.43	222.44	14.65	222.44	14.65

Table 7a

MAP Growth Model Results- Elementary Less than 95th Percentile Sample

	Elementary Mathematics		Elementary Reading	
	Model 1	Model 2	Model 1	Model 2
Fixed Effects				
Intercept	193.29 (.04)***	193.29 (.04)***	195.11 (.05)***	195.11 (.05)***
Time	0.87 (.001)***	0.87 (.001)***	0.64 (.001)***	0.64 (.001)***
Male	0.33 (.05)***	0.33 (>05)***	-2.70 (.05)***	-2,70 (.05)***
Title I	-1.84 (.02)***	-1.84 (.02)***	-2.54 (.02)***	-2.54 (.02)***
Minority	-5.70 (.05)***	-5.70 (.05)***	-7.28 (.05)***	-7.28 (.05)***
Time X Male	0.01 (.001)***	0.01 (.001)***	0.004 (.001)***	0.004 (.001)***
Time X Title I	0.52 (.002)***	0.52 (.002)***	0.53 (.002)***	0.53 (.002)***
Time X Minority	-0.05 (.001)***	-0.05 (.001)***	0.02 (.001)***	0.02 (.001)***
Variance Components				
Intercept	151.35	151.34	191.28	191.28
Residual	33.66	33.66	45.16	45.16
ICC	0.82	0.82	0.81	0.81

Note: * p < .05, ** p < .01, *** p < .001

Table 7b

MAP Growth Model Results- Middle School Less than 95th Percentile Sample

	Middle School Mathematics		Middle School Reading	
	Model 1	Model 2	Model 1	Model 2
Fixed Effects				
Intercept	218.53 (.05)***	218.54 (.05)***	215.50 (.05)***	215.49 (.05)***
Time	0.56 (.001)***	0.56 (.001)***	0.35 (.05)***	0.35 (.001)***
Male	0.03 (.06)	0.03 (.06)	-2.82 (.05)***	-2.82 (.05)***
Title I	-0.48 (.02)***	-0.49 (.02)***	-0.53 (.02)***	-0.53 (.02)***
Minority	-7.36 (.06)***	-7.36 (.06)***	-6.55 (>05)***	-6.55 (.05)***
Time X Male	-0.04 (.001)***	-0.04 (.001)***	-0.04 (.001)***	-0.04 (.001)***
Time X Title I	0.29 *.002)***	0.29 (.002)***	0.20 (.002)***	0.20 (.002)***
Time X Minority	-0.03 (.001)***	-0.03 (.001)***	0.04 (.001)***	0.04 (.001)***
Variance Components				
Intercept	234.87	234.72	185.99	185.99
Residual	32.75	32.72	41.30	41.30
ICC	0.88	0.88	0.82	0.82

Note: * p < .05, ** p < .01, *** p < .001

Greater Than or Equal to 95th Percentile Achievement

This sample of students had higher levels of average overall achievement, as was expected since this sample was specifically selected for their higher percentile rank. The overall means and standard deviations were as follows, 236.83 (*SD* = 13.42), 229.46 (*SD* = 8.24), 260.39 (*SD* = 9.74), and 247.54 (*SD* = 5.47), for elementary math (n = 34,308) and reading (n = 26,735) and middle school math (n = 29,468) and reading (n = 17,607), respectively. Table 8 shows descriptive statistics for both the uncensored and censored test scores broken down by semester. The ICCs were lower for the high ability samples (ranging from .32 to .61 for the uncensored models and approximately <0.001 to .21 for the artificially censored models). The ICCs decreased because as more observations were censored, the variability between students

decreased (sometimes disappeared entirely) while the variability within students still existed across time points.

Table 8

MAP Growth Mean Achievement Disaggregated by Time- 95th Percentile

	Elementary Mathematics				Elementary Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall '16	215.54	6.08	212.77	1.83	218.55	4.75	217.27	2.58
Winter '16	223.84	6.16	220.81	1.67	224.49	4.15	223.10	1.98
Spring '17	230.47	6.33	225.70	0.68	228.49	4.24	226.16	1.16
Fall '17	229.32	6.52	225.86	1.79	228.36	3.98	226.69	1.64
Winter '17	236.76	6.73	232.97	1.56	232.65	3.46	231.06	1.27
Spring '18	243.00	6.92	237.69	0.87	235.59	3.76	232.77	0.71
Fall '18	240.52	6.78	237.84	2.32	235.41	3.88	232.68	0.77
Winter '18	247.52	7.10	244.19	2.18	239.16	3.92	235.88	0.38
Spring '19	254.09	7.15	249.23	2.00	241.77	4.29	236.99	0.21
	Middle School Mathematics				Middle School Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall '16	247.29	5.92	244.55	1.76	241.26	3.72	240.60	2.40
Winter '16	252.46	5.79	249.75	1.56	243.86	3.55	242.88	1.83
Spring '17	257.44	6.15	253.53	0.82	245.70	3.65	244.41	1.55
Fall '17	255.11	6.34	252.36	1.88	245.28	3.62	243.78	1.33
Winter '17	259.58	6.36	256.63	1.69	247.73	3.76	245.47	0.79
Spring '18	264.52	7.19	260.19	1.24	249.74	4.00	246.74	0.49
Fall '18	263.09	7.13	256.92	0.27	250.35	4.13	248.67	1.44
Winter '18	266.83	7.19	259.97	0.20	252.38	4.21	250.19	1.08
Spring '19	270.87	7.37	262.00	0.10	254.05	4.24	251.47	0.81

Table 9 shows the results of each model for this sample. In this sample, the artificially censored elementary cohort estimates decreased by .07 and .08 for math and reading, respectively. The Tobit model increased estimates by .03 units for math and .05 units for reading. Ultimately, this means that the Tobit models were able to make up for 42.9% of the elementary math censoring, and 62.5% for elementary reading. Therefore, the Tobit models for elementary reading were better able to reproduce the uncensored estimates. The elementary math cohort had 9.48% of observations censored, and the elementary reading cohort had 7.28% of observations censored at the ceiling.

Table 9a

MAP Growth Model Results- Elementary 95th Percentile Sample

	Elementary Mathematics			Elementary Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	214.37 (.09)***	213.74 (.03)***	213.27 (.03)***	219.84 (.05)***	219.48 (.03)***	218.86 (.03)***
Time	1.13 (.003)***	1.06 (.001)***	1.09 (.001)***	0.64 (.003)***	0.56 (.001)***	0.61 (.002)***
Male	1.30 (.09)***	0.24 (.03)***	0.25 (.03)***	0.27 (.06)***	0.14 (.03)***	0.15 (.03)***
Title I	-1.77 (.08)***	-1.54 (.03)***	-1.14 (.03)***	-2.72 (.06)***	-3.12 (.03)***	-2.58 (.03)***
Minority	0.70 (.09)***	0.18 (.03)***	0.18 (.03)***	-0.14 (.06)*	-0.18 (>03)***	-0.13 (.03)***
Time X Male	-0.03 (.003)***	-0.01 (.001)***	-0.01 (.001)***	-0.01 (.003)*	-0.002 (.002)	-0.003 (.002)*
Time X Title I	0.60 (.01)***	0.44 (.004)***	0.42 (.004)***	0.54 (.01)***	0.51 (.004)***	0.48 (.004)***
Time X Minority	0.02 (.003)***	-0.001 (.001)	0.001 (.002)	0.01 (.003)	0.01 (.002)***	0.01 (.002)***
Variance Components						
Intercept	19.95	1.16	1.25	6.17	0.95	0.92
Residual	24.27	4.57	5.24	11.92	3.62	3.65
ICC	0.45	0.20	0.19	0.34	0.21	0.20

Note. * p < .05, ** p < .01, *** p < .001

The middle school math cohort had 13.98% of observations censored, and 7.18% for middle school reading. In the middle school cohorts, artificially censored estimates decreased by .28 units and .06 units for math and reading, respectively. The Tobit models were able to increase the censored estimates by .06 units for mathematics and .02 units for reading. This means that the Tobit models were able to make up for 21% of the censoring for middle school math and 33% for middle school reading. Potential explanations for these results will be discussed in the next chapter.

Table 9b

MAP Growth Model Results- Middle School 95th Percentile Sample

	Middle School Mathematics			Middle School Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	244.39 (.08)***	246.95 (.03)***	246.14 (.03)***	240.44 (.06)***	240.01 (.02)***	239.84 (.03)***
Time	0.72 (.003)***	0.44 (.001)***	0.50 (.001)***	0.39 (.002)***	0.33 (.001)***	0.35 (.001)***
Male	1.14 (.08)***	0.06 (.03)*	0.12 (.03)***	0.18 (.07)**	0.11 (.03)***	0.13 (.03)***
Title I	-1.19 (.07)***	-3.29 (.03)***	-2.72 (.04)***	-0.45 (.07)***	0.01 (.03)	0.16 (.03)***
Minority	0.65 (.09)***	0.15 (.03)***	0.16 (.03)***	0.06 (.07)	0.09 (.03)**	0.09 (.03)**
Time X Male	-0.01 (.003)***	0.002 (.001)	-0.001 (.001)	-0.02 (.003)***	-0.01 (.001)***	-0.01 (.001)***
Time X Title I	0.41 (.01)***	0.48 (.004)***	0.46 (.01)***	0.17 (.01)***	0.13 (.004)***	0.12 (.004)***
Time X Minority	0.05 (.003)***	-0.003 (.001)*	-0.001 (.002)	0.001 (.003)	-0.003 (.001)*	0.12 (.004)
Variance Components						
Intercept	24.42	< 0.001	< 0.001	4.60	0.52	0.55
Residual	15.36	4.39	4.86	9.50	1.93	2.09
ICC	0.61	< 0.001	< 0.001	0.32	0.21	0.21

Note. * p < .05, ** p < .01, *** p < .001

Greater Than or Equal to 98th Percentile Achievement

This sample of students had the highest overall test scores across all time points. Mean elementary mathematics (n = 15,476) and reading (n = 7,661) scores were 243.29 (SD = 13.41) and 234.55 (*SD* = 8.25), respectively. Mean middle school mathematics (n = 11,402) and reading (n = 4,440) scores were 267.55 (*SD* = 10.36) and 253.72 (*SD* = 5.46), respectively. Table 10 contains the descriptive statistics for this sample at each time point.

Table 10

MAP Growth Mean Achievement Disaggregated by Time- 98th Percentile

Year	Elementary Mathematics				Elementary Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall '16	221.26	6.62	213.88	1.20	224.54	4.64	219.83	1.61
Winter '16	229.59	6.31	221.96	0.61	230.25	3.77	224.89	1.08
Spring '17	236.04	6.20	225.99	0.30	234.25	3.56	227.00	0.08
Fall '17	234.89	6.37	226.95	0.73	234.19	3.45	227.94	0.85
Winter '17	242.49	6.58	233.98	0.36	238.42	3.10	231.96	0.69
Spring '18	249.25	6.84	237.97	0.51	241.82	3.44	232.99	0.13
Fall '18	246.66	6.85	239.96	0.78	241.60	3.69	232.96	1.03
Winter '18	253.80	6.74	245.98	0.38	245.56	3.72	236.00	0.00
Spring '19	260.02	6.54	249.98	0.32	248.60	4.39	237.00	0.00
	Middle School Mathematics				Middle School Reading			
	Uncensored		Censored		Uncensored		Censored	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fall '16	253.17	5.77	245.00	0.01	247.17	3.07	243.99	0.11
Winter '16	258.90	6.07	250.00	0.00	249.65	3.03	245.00	0.00
Spring '17	264.71	6.63	253.00	0.00	251.56	3.12	246.00	0.00
Fall '17	261.68	6.52	253.00	0.10	251.36	3.18	245.00	0.00
Winter '17	266.65	6.67	257.00	0.00	253.74	3.48	246.00	0.00
Spring '18	272.04	7.17	260.00	0.00	255.90	3.46	247.00	0.00
Fall '18	270.64	7.38	256.00	0.08	257.08	4.08	250.00	0.06
Winter '18	274.58	7.22	259.00	0.16	258.98	4.12	251.00	0.00
Spring '19	278.60	7.10	261.00	0.00	260.92	4.14	252.00	0.00

The ICCs for this sample ranged from .18 to .42 for the uncensored models and < 0.001to .08 for the censored models. At the 98th percentile, 15.8% of observations were censored for elementary math, 6.34% for elementary reading, 14.68% for middle school math, and 9.25% for middle school reading. The results of all models for each cohort can be found in Table 11.

Table 11a

MAP Growth Model Results- Elementary 98th Percentile Sample

	Elementary Mathematics			Elementary Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	220.07 (.14)***	215.02 (.03)***	214.30 (.04)***	225.28 (.10)***	221.38 (.04)***	220.99 (.04)***
Time	1.15 (.01)***	1.07 (.001)***	1.12 (.002)***	0.68 (.01)***	0.50 (.002)***	0.54 (.002)***
Male	1.30 (.14)***	0.03 (.03)	0.10 (.04)**	0.27 (.11)*	0.11 (.04)**	0.12 (.04)**
Title I	-2.04 (.13)***	-1.24 (.04)***	-0.71 (.04)***	-2.05 (.12)***	-2.20 (.05)***	-1.92 (.05)***
Minority	0.80 (.13)***	0.01 (.03)	0.07 (.03)*	0.01 (.12)	-0.13 (.04)**	-0.11 (.05)*
Time X Male	-0.03 (.01)***	-0.001 (.004)***	-0.01 (.002)***	-0.003 (.01)	-.001 (.01)	-0.003 (.003)
Time X Title I	0.60 (.02)***	0.35 (.004)***	0.33 (.01)***	0.48 (.01)***	0.38 (>01)***	0.37 (.01)***
Time X Minority	0.02 (.01)***	-0.001 (.001)	-0.004 (.002)*	0.003 (.01)	0.01 (.003)*	0.01 (.003)*
Variance Components						
Intercept	17.74	0.19	0.16	4.88	0.12	0.13
Residual	25.43	2.25	2.67	11.59	2.27	2.24
ICC	0.41	0.08	0.06	0.30	0.05	0.05

Note. * p < .05, ** p < .01, *** p < .001

Table 11b

MAP Growth Model Results- Middle School 98th Percentile Sample

	Middle School Mathematics			Middle School Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	250.07 (.14)***	248.01 (.04)***	247.24 (.05)***	246.40 (.12)***	242.53 (.03)***	242.38 (.03)***
Time	0.77 (.01)***	0.41 (.002)***	0.46 (.002)***	0.43 (.01)***	0.27 (.001)***	0.28 (.002)***
Male	1.07 (.14)***	-0.09 (.04)*	-0.03 (.05)	-0.08 (.14)	0.11 (.03)**	0.12 (.04)**
Title I	-1.28 (.14)***	-3.22 (.05)***	-2.76 (.06)***	-.018 (.17)	1.10 (.04)***	1.24 (.05)***
Minority	1.11 (.14)***	-0.11 (.04)**	-0.09 (.05)*	0.24 (.15)	0.11 (.04)**	0.12 (.04)**
Time X Male	-0.002 (.01)	0.01 (.002)**	0.004 (.002)	-0.01 (.01)*	-0.004 (.002)*	-0.01 (.002)**
Time X Title I	0.44 (.02)***	0.44 (.01)***	0.45 (.01)***	0.15 (.02)***	-0.01 (.01)*	-0.02 (.01)***
Time X Minority	0.05 (.01)***	0.01 (.002)**	0.01 (.002)*	-0.01 (.01)	-0.004 (.002)*	-0.004 (>002)
Variance Components						
Intercept	11.84	< 0.001	< 0.001	2.21	< 0.001	< 0.001
Residual	16.57	3.85	4.24	9.93	0.84	0.97
ICC	0.42	< 0.001	< 0.001	0.18	< 0.001	< 0.001

Note. * p < .05, ** p < .01, *** p < .001

The artificially censored elementary cohort estimates decreased by .08 units and .18 units for math and reading, respectively. The multilevel Tobit models were able to increase the censored estimates by .05 units for mathematics and .04 units for reading; meaning 62.5% of

censoring was corrected for in mathematics and 22% in reading. The middle school cohort of the 98th percentile MAP Growth sample had estimates decrease by .36 units for mathematics and .16 units for reading. Subsequently, the Tobit models increased estimates by .05 units and approximately .01 units, respectively; approximately 14% of censoring was corrected for in mathematics and 12% for reading. Given the 95th percentile results combined with these, it is possible that both the percentage of censored observations and the severity of censorship have an impact on the performance of the Tobit model. This will be discussed in further detail in the next chapter.

Part Two: NC EOG

The second part of this study intended to answer the following research question: Are growth estimates from Tobit regression significantly different from those obtained using traditional multilevel models on naturally censored data? To answer this question, up to three models were run for each sample and cohort of data (whole group sample and AG only sample for elementary math and reading cohorts and middle school math and reading cohorts). Model 1 used NC EOG test scores, which were demonstrated to be naturally censored, as the outcome and time after testing (in years) and demographic variables as the predictors, with clustering occurring at the student level. Model 2 was the multilevel Tobit regression with the same outcome and predictor variables and the same clustering; the censor point for the model was set at the test ceiling despite censoring occurring at up to three time points across the data. Model 3 was another multilevel Tobit regression using the same variables, however the ceiling was lowered to the 95th percentile in order to account for censoring that could have been missed due to naturally occurring measurement error. Table 12 shows the percentage of censored observations at both the ceiling and the 95th percentile ceiling.

Table 12

NC EOG Percentage of Observations Censored at the Ceiling and at the 95th Percentile

Sample and Cohort	% (Ceiling)	% (95th Percentile)
Whole Group		
Elementary Mathematics	0.08	2.28
Elementary Reading	0.04	2.90
Middle School Mathematics	0.12	2.88
Middle School Reading	0.04	3.47
AG Sample		
Elementary Mathematics	0.58	12.19
Elementary Reading	0.31	16.09
Middle School Mathematics	0.71	14.70
Middle School Reading	0.28	16.68
Non-AG Sample		
Elementary Mathematics	0.06	1.86
Elementary Reading	0.03	2.35
Middle School Mathematics	0.05	1.51
Middle School Reading	0.02	2.02

Note. AG = Academically Gifted

Overall, across samples and cohorts the results demonstrated similar trends as the MAP Growth portion of the study. One major difference, however, was in Part One the higher ability students grew larger amounts, whereas the estimates from this part of the study showed that higher ability students grew less. Potential explanations and implications of these results will be discussed in the following chapter.

Whole Group Sample

Average achievement scores across all time points for the whole group sample for each cohort were 352.02 (*SD* = 9.92), 346.75 (*SD* = 10.53), 360.94 (*SD* = 8.68), and 357.79 (*SD* = 8.52) for elementary math and reading and middle school math and reading, respectively. Similar to the MAP Growth assessment data, students scored higher in math than reading for both age group cohorts. Mean achievement disaggregated by time for each cohort is shown in Table 13. It is important to remember that the EOG estimates represented yearly growth estimates, unlike the

monthly growth rates for the MAP Growth estimates. Additionally, the scales of the MAP Growth assessment and the EOG assessment differ and therefore the estimates (and thus the changes in estimates) were not directly comparable.

Table 13

NC EOG Mean Achievement Disaggregated by Time

Year	Elementary		Middle School	
	Mathematics	Reading	Mathematics	Reading
2009	345.90 (8.84)	340.74 (10.83)	357.70 (8.42)	354.56 (8.62)
2010	352.49 (8.63)	347.27 (8.96)	361.08 (8.55)	358.01 (8.06)
2011	357.78 (8.85)	352.38 (8.14)	364.01 (7.87)	360.81 (7.67)

The results of the full models for this sample are in Table 14. The resulting changes from Model 1 to Model 2 were negligible; this is likely due to such small percentages of censored observations. These results add to the patterns that emerged in the first part of this study.

Table 14a

NC EOG Model Results- Elementary Whole Sample

	Elementary Mathematics			Elementary Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	343.13 (.05)***	343.13 (.05)***	343.07 (.05)***	340.57 (.06)***	340.56 (.06)***	340.48 (.06)***
Time	5.62 (.01)***	5.63 (.01)***	5.67 (.02)***	4.84 (.02)***	4.84 (.02)***	4.91 (.02)***
Male	1.10 (.05)***	1.10 (.05)***	1.10 (.05)***	-0.94 (.06)***	-0.94 (.06)***	-0.93 (.06)***
Title I	-2.74 (.05)***	-2.74 (.05)***	-2.72 (.06)***	-3.44 (.06)***	-3.44 (.06)***	-3.41 (.06)***
Minority	-5.02 (.06)***	-5.02 (.06)***	-4.99 (.06)***	-7.09 (.06)***	-7.08 (.06)***	-7.04 (.06)***
Time X Male	-0.07 (.02)***	-0.07 (.02)***	-0.06 (.02)***	0.19 (.02)***	0.19 (.02)***	0.19 (.02)***
Time X Title I	-0.02 (.02)	-0.02 (.02)	-0.04 (.02)*	0.27 (.02)***	0.27 (.02)***	0.24 (.02)***
Time X Minority	0.38 (.05)***	0.38 (.02)***	0.36 (.05)***	0.98 (.02)***	0.98 (.02)***	0.95 (.02)***
Variance Components						
Intercept	57.41	57.45	58.05	65.97	65.99	67.15
Residual	12.84	12.85	12.99	16.4	16.41	16.55
ICC	0.82	0.82	0.82	0.80	0.80	0.80

Note. * p < .05, ** p < .01, *** p < .001

Table 14b

NC EOG Model Results- Middle Schools Whole Sample

	Middle School Mathematics			Middle School Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	356.82 (.04)***	356.80 (.04)***	356.77 (.05)***	354.74 (.04)***	354.74 (.04)***	354.71 (.04)***
Time	2.96 (.01)***	2.96 (.01)***	2.99 (.01)***	2.77 (.01)***	2.77 (.01)***	2.81 (.01)***
Male	0.39 (.05)***	0.40 (.05)***	0.40 (.05)***	-1.43 (.05)***	-1.43 (.05)***	-1.43 (.05)***
Title I	-2.47 (.07)***	-2.46 (.07)***	-2.47 (.08)***	-2.17 (.07)***	-2.17 (.07)***	-2.16 (.07)***
Minority	-5.44 (.05)***	-5.43 (.05)***	-5.41 (.05)***	-5.52 (.05)***	-5.52 (.05)***	-5.50 (>05)***
Time X Male	-0.23 (.01)***	-0.23 (.01)***	-0.23 (.01)***	0.27 (.01)***	0.27 (.01)***	0.27 (.02)***
Time X Title I	0.31 (.02)***	0.31 (.02)***	0.31 (.02)***	0.10 (.02)***	0.10 (.02)***	0.10 (.02)***
Time X Minority	0.53 (.02)***	0.53 (.02)***	0.51 (.02)***	0.33 (.02)***	0.33 (.02)***	0.30 (.02)***
Variance Components						
Intercept	53.98	54.06	54.63	50.46	50.48	51.36
Residual	10.76	10.79	10.82	10.86	10.87	10.96
ICC	0.83	0.83	0.83	0.82	0.82	0.82

Note. * p < .05, ** p < .01, *** p < .001

Academically Gifted Sample

As previously explained, this sample of students was selected by identifying those who were labeled AG by the state of North Carolina. Although high test scores are one potential path to identification of gifted status, it is not the only path and therefore this population of students did not have exclusively high test scores. That said, average achievement across all cohorts and subjects was still higher than the whole group sample: 362.86 (*SD* = 7.16), 357.92 (*SD* = 6.82), 371.08 (*SD* = 6.17), and 367.07 (*SD* = 5.50) for elementary mathematics (n = 4,786) and reading (n = 4,703) and middle school mathematics (n = 11,412) and reading (n = 10,784), respectively. Mean achievement disaggregated by time is shown in Table 15.

Table 15

NC EOG Mean Achievement Disaggregated by Time

Year	Elementary		Middle School	
	Mathematics	Reading	Mathematics	Reading
2009	357.39 (5.76)	354.09 (6.63)	368.36 (5.74)	364.97 (5.47)
2010	363.35 (5.35)	358.13 (5.83)	371.62 (5.83)	366.86 (5.00)
2011	368.43 (5.52)	361.98 (5.45)	373.44 (5.84)	369.52 (5.03)

The AG sample using the EOG data sets were considered naturally censored, therefore it was not possible to determine the true amount of censorship existing in the estimates. However, the percentages of censored observations at the ceiling of each grade is known. These proportions of observations censored at the ceiling were most similar to the MAP Growth whole group sample. Therefore, it was unsurprising that the Tobit multilevel estimates had only small increases (see Table 16 for full model results). However, the proportions of observations censored when the ceiling was lowered to the 95th percentile was most similar to the MAP Growth 98th percentile sample. Since these censor percentages were larger, it was reasonable that the estimates increased more for these models. The largest improvements from Model 1 to Model 3 existed for elementary math and reading. The larger changes in estimates using the lower-ceiling model demonstrated the importance of considering the impact of random measurement error on high-ability students' scores.

Table 16a

NC EOG Model Results- Elementary AG Sample

	Elementary Mathematics			Elementary Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	352.47 (.17)***	352.45 (.17)***	352.28 (.18)***	351.78 (.19)***	351.76 (.19)***	351.69 (.19)***
Time	5.60 (.06)***	5.61 (.06)***	5.76 (.07)***	3.79 (.07)***	3.80 (.07)***	3.93 (.08)***
Male	1.24 (.20)***	1.23 (.20)***	1.22 (.21)***	-0.22 (.22)	-0.22 (.22)	-0.27 (.23)
Title I	-1.05 (.21)***	-1.03 (.21)***	-1.01 (.21)***	-1.56 (,23)***	-1.56 (,23)***	-1.52 (.24)***
Minority	-2.69 (.22)***	-2.68 (.22)***	-2.61 (.21)***	-3.69 (.25)***	-3.68 (.25)***	-3.68 (.26)***
Time X Male	-0.20 (.07)**	-0.19 (.08)*	-0.18 (.08)*	-0.04 (.09)	-0.04 (.09)	-0.02 (.09)
Time X Title I	-0.07 (.08)	-0.08 (.08)	-0.11 (.08)	0.16 (.09)	0.16 (.09)	0.10 (.10)
Time X Minority	0.28 (.08)**	0.27 (.17)***	0.21 (.09)*	0.48 (.10)***	0.47 (.10)***	0.44 (.10)***
Variance Components						
Intercept	16.99	17.11	18.17	18.43	18.49	20.13
Residual	12.34	12.47	13.15	15.99	16.08	16.82
ICC	0.58	0.58	0.58	0.54	0.53	0.54

Note. * p < .05, ** p < .01, *** p < .001

Table 16b

NC EOG Model Results- Middle School AG Sample

	Middle School Mathematics			Middle School Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	366.23 (.10)***	366.21 (.10)***	366.14 (.10)***	363.55 (.09)***	363.55 (.09)***	363.53 (.10)***
Time	2.66 (.03)***	2.68 (.03)***	2.74 (.04)***	2.18 (.04)***	2.19 (.04)***	2.26 (.04)***
Male	0.80 (.13)***	0.80 (.013)***	0.82 (.13)***	-0.83 (.12)***	-0.83 (.13)***	-0.83 (.13)***
Title I	-2.37 (.22)***	-2.35 (.22)***	-2.38 (.23)***	-2.31 (.21)***	-2.30 (.21)***	-2.29 (.22)***
Minority	-1.53 (.15)***	-1.54 (.15)***	-1.46 (.15)***	-1.64 (.12)***	-1.64 (.15)***	1.64 (.16)***
Time X Male	-0.36 (.04)***	-0.36 (.04)***	-0.37 (.05)***	0.10 (.05)*	0.10 (.05)*	0.09 (.05)
Time X Title I	0.21 (.07)***	0.20 (.08)**	0.21 (.08)**	0.20 (.08)*	0.20 (.08)*	0.17 (.08)*
Time X Minority	0.19 (.05)***	0.20 (.05)***	0.14 (.05)**	0.10 (.06)	0.10 (.06)	0.08 (.06)
Variance Components						
Intercept	23.00	23.18	24.01	14.36	14.41	15.45
Residual	10.16	10.29	10.34	11.69	11.76	12.23
ICC	0.69	0.69	0.699	0.55	0.55	0.56

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Non-AG Sample

This sample was primarily selected to clearly demonstrate the differences in growth patterns between AG and non-AG students. Since students were not identified as AG simply through high test scores, there was still some censoring at both the ceiling and 95th percentile for this sample. These percentages were similar, however, to the whole group sample. Additionally, sample sizes were much larger than the AG sample (n = 125,237; n = 124,497; n = 108,834; n = 108,900 for elementary mathematics and reading, and middle school mathematics and reading, respectively). There were negligible changes from Model 1 to Model 2, however some significant changes occurred when the ceiling was lowered. The largest improvements were

made for elementary reading. See Table 17 for average achievement disaggregated by time and Table 18 for full results for this sample.

Table 17

NC EOG Mean Achievement Disaggregated by Time- Non-AG Sample

Year	Elementary		Middle School	
	Mathematics	Reading	Mathematics	Reading
2009	345.40 (8.60)	340.17 (10.60)	356.41 (7.5)	353.38 (8.09)
2010	352.04 (8.44)	346.83 (8.78)	359.86 (7.95)	357.04 (7.73)
2011	357.36 (8.26)	352.02 (7.99)	362.96 (7.34)	359.88 (7.32)

Importantly, by separating the AG and non-AG students, it was possible to see that across all subjects and cohorts, the non-AG students demonstrated more growth than their AG peers. These results directly contradicted the growth patterns from the MAP Growth portion of the study and will be discussed further in the following chapter.

Table 18a

NC EOG Model Results- Elementary Non-AG Sample

	Elementary Mathematics			Elementary Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	342.61 (.05)***	342.60 (.05)***	342.56 (.02)***	339.85 (.06)***	339.85 (.06)***	339.78 (.06)***
Time	5.64 (.02)***	5.64 (.02)***	5.68 (.02)***	4.93 (.02)***	4.93 (.02)***	4.99 (.02)***
Male	0.99 (.05)***	0.99 (.05)***	0.99 (.05)***	-1.01 (.06)***	-1.01 (.06)***	-1.00 (.06)***
Title I	-2.57 (.06)***	-2.56 (.06)***	-2.55 (.06)***	-3.22 (.06)***	-3.21 (.06)***	-3.20 (.06)***
Minority	-4.86 (.06)***	-4.86 (.06)***	-4.83 (.06)***	-6.84 (.06)***	-6.84 (.06)***	-6.81 (.06)***
Time X Male	-0.07 (.02)***	-0.07 (.02)***	-0.06 (.02)***	0.19 (.02)***	0.19 (.02)***	0.19 (.02)***
Time X Title I	-0.02 (.02)	-0.02 (.02)	-0.04 (.02)*	0.24 (.02)***	0.24 (.02)***	0.22 (.02)***
Time X Minority	0.38 (.02)***	0.38 (.02)***	0.36 (.02)***	0.95 (.02)***	0.95 (.02)***	0.93 (.02)***
Variance Components						
Intercept	54.53	54.57	55.04	63.27	63.29	64.18
Residual	12.86	12.88	12.98	16.34	16.35	16.46
ICC	0.81	0.81	0.81	0.79	0.79	0.80

Note. * p < .05, ** p < .01, *** p < .001

Table 18b

NC EOG Model Results- Middle School Non-AG Sample

	Middle School Mathematics			Middle School Reading		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Fixed Effects						
Intercept	355.39 (.05)***	355.39 (.05)***	355.36 (.05)***	353.33 (.05)***	353.32 (.05)***	353.31 (.05)***
Time	3.03 (.01)***	3.04 (.01)***	3.05 (.01)***	2.90 (.01)***	2.90 (.01)***	2.92 (.01)***
Male	0.25 (.05)***	0.25 (.05)***	0.26 (.05)***	-1.38 (.05)***	-1.38 (.05)***	-1.38 (.05)***
Title I	-2.04 (.07)***	-2.03 (.07)***	-2.04 (.07)***	-1.85 (.07)***	-1.85 (.07)***	-1.85 (.07)***
Minority	-4.84 (.05)***	-4.84 (.05)***	-4.82 (.05)***	-4.85 (.05)***	-4.85 (.05)***	-4.84 (.05)***
Time X Male	-0.22 (.02)***	-0.22 (.02)***	-0.22 (.02)***	0.27 (.02)***	0.27 (.02)***	0.26 (.02)***
Time X Title I	0.29 (.02)***	0.29 (.02)***	0.29 (.02)***	0.07 (.02)**	0.07 (.02)**	0.07 (.02)**
Time X Minority	0.51 (.02)***	0.51 (.02)***	0.50 (.02)***	0.26 (.02)***	0.26 (.02)***	0.24 (.02)***
Variance Components						
Intercept	45.81	45.84	46.08	45.69	45.7	46.18
Residual	10.81	10.82	10.84	10.70	10.70	10.75
ICC	0.81	0.81	0.81	0.81	0.81	0.81

Note. * p < .05, ** p < .01, *** p < .001

Summary of Results

This study sought to answer the following research questions:

1. Can Tobit Regression be used to replicate uncensored growth estimates on artificially censored data?

    1a. At what level of achievement is Tobit Regression preferable over traditional modeling procedures?

2. Are growth estimates from Tobit Regression significantly different from those obtained using traditional models on naturally censored data?

Ultimately, the Tobit models using artificially censored data were able to come close to replicating the uncensored growth estimates under certain conditions. The results indicated that Tobit regression was necessary when examining homogeneous groups of high ability students. Finally, the Tobit regression models were able to increase the growth estimates for high ability

students using naturally censored data. The degree to which the models increased and under which conditions the increases existed will be discussed in the following chapter.

CHAPTER 5: DISCUSSION

Discussion

This study sought to determine if Tobit regression could be used to model high-ability student growth when the measure of achievement had ceiling effects. The trends presented in Chapter 4 made evident potential conditions under which the Tobit model was best able to increase previously censored estimates. First, the proportion of censored observations impacted both the necessity of using the Tobit model and the corrections made by the Tobit model. Additionally, the severity of the censor impacted the Tobit model's corrections. Although it was impossible to draw official demarcations for the degree of censoring and the proportions of censored observations from the results of this study alone, this study has demonstrated that substantial gains in growth estimates could be produced when approximately 5% to 15% of observations were censored. Table 19 contains a summary of the growth coefficients for all models under all conditions.

Table 19

Summary Table of Growth Coefficients

	MAP Growth Coefficients											
	Elementary Mathematics			Elementary Reading			Middle School Mathematics			Middle School Reading		
	1	2	3	1	2	3	1	2	3	1	2	3
Whole Sample	0.89	0.89	0.89	0.63	0.63	0.63	0.58	0.56	0.57	0.35	0.35	0.35
Less Than 95th Percentile	0.87	0.87		0.64	0.64		0.56	0.56		0.35	0.35	
95th Percentile	1.13	1.06	1.09	0.64	0.56	0.61	0.72	0.44	0.50	0.39	0.33	0.35
98th Percentile	1.15	1.07	1.12	0.68	0.50	0.54	0.77	0.41	0.46	0.43	0.27	0.28
	NC EOG Growth Coefficients											
	Elementary Mathematics			Elementary Reading			Middle School Mathematics			Middle School Reading		
	1	2	3	1	2	3	1	2	3	1	2	3
Whole Sample	5.62	5.63	5.67	4.84	4.84	4.91	2.96	2.96	2.99	2.77	2.77	2.81
AG	5.60	5.61	5.76	3.79	3.8	3.93	2.66	2.68	2.74	2.18	2.19	2.26
Non-AG	5.64	5.64	5.68	4.93	4.93	4.99	3.03	3.04	3.05	2.90	2.90	2.92

Note. AG = Academically Gifted; 1 = Model 1; 2 = Model 2; 3 = Model 3.

Additionally, the results highlighted patterns in growth when compared by subject, cohort, and ability level. Specifically, for both assessments students showed greater growth in math than reading. Additionally, the elementary cohort showed greater growth than the middle school cohort. Finally, the MAP Growth data sets demonstrated that higher ability students made higher growth than their peers, whereas the EOG data sets demonstrated that higher ability students made lower growth than their peers.

Proportion of Censored Observations

The results of this study demonstrated that increases in growth estimates were produced when approximately 5% to 15% of observations were censored. Specifically, the largest corrections were made for the MAP Growth 95th percentile sample of elementary math and reading, the MAP Growth 98th percentile sample of elementary math, and the NC EOG AG sample for elementary math and reading. Previous studies have examined the impact of the amount of censoring on various aspects of performance for Tobit models. Leiker (2009) used simulations to study the effect of sample size on the performance of Tobit models while simultaneously manipulating the percentage of censored observations. The author concluded that when the proportion of censored observations neared 50%, the estimates were less accurate. Arabmazar and Schmidt (1981) explored the impact of violations of the homoscedasticity assumption and found that when heteroscedasticity increased in conjunction with the proportion of censored observations, the Tobit estimates were more inconsistent. Paarsch (1984) also found issues with the estimates produced by the Tobit model when censoring approached 50% of the sample. Finally, Wang et al. (2008) experimented with various percentages of censored observations and found that as those percentages reached 40%, bias in the estimates was larger.

Severity of Censoring

Another potential condition that had an impact on the performance of the Tobit models was the severity of censoring (i.e., the degree to which estimates decreased when using the censored outcome variable). Specifically, the estimates using the MAP Growth 95th percentile middle school math sample decreased by approximately .3 units when using the censored outcome variable versus the uncensored outcome variable. The Tobit model was able to increase estimates by approximately .06 units, or approximately a 21% improvement. Similarly, the estimates using the MAP Growth 98th percentile middle school math decreased by nearly .4 units (the largest decrease of any MAP Growth models), and the Tobit model only increased the estimates by .05 units (or approximately a 14% improvement). In comparison, the samples with the biggest improvements (change increase/change decrease) were the MAP Growth 95th percentile elementary math and reading samples which only decreased from Model 1 to Model 2 by .07 and .08, respectively.

Despite similar proportions of censoring across most of the MAP Growth models, the Tobit models' ability to replicate the uncensored estimates was worse when the sample was more severely censored. Of all of the research exploring the methodological soundness and conditions under which the Tobit model best performs, no studies could be found that explored the severity of censoring as a potentially important condition. This is most likely due to the additional lack of research using Tobit models with ceiling effects rather than floor effects.

Subject and Cohort Growth Trends

For both the MAP Growth data and NC EOG data, elementary growth estimates were higher than middle school estimates. This trend has been found in nearly all studies examining growth in student achievement (e.g., Baumert et al., 2012; Cameron et al., 2015; Lee, 2010; Scamacca et al., (2020). Cameron et al. (2015) explained that this trend is likely explained due to

the initial rapid acquisition of knowledge and skills in the younger grades which is then followed by a slower knowledge building process in the older grades. Since this study did not include a quadratic time variable to determine the pace of the growth trends, it can be loosely concluded that the lower growth in middle school is further evidence of growth trends decreasing over time.

Across all samples and grade level cohorts, growth estimates were lower in reading than mathematics. The previously mentioned research demonstrated the decrease in growth trends as students aged; additionally Scammacca et al. (2020) found that the pace of growth in reading slowed more than that of math although both decreased. Rambo-Hernandez et al. (2021) also found lower growth in reading than mathematics. Specifically, the authors found relatively consistent growth when compared across achievement levels for mathematics, likely due to the subject more naturally increasing in difficulty over time (therefore students were more likely to be within their ZPD). Lee (2010) posited that since math concepts naturally become more complex they also take longer to master, therefore resulting in slowed growth, but not as slow as for reading. Rambo-Hernandez et al. (2021) also concluded that students with higher initial achievement in reading had significantly slower rates of growth, likely due to content being outside of their ZPD (i.e. the content was too easy, therefore learning could not occur).

Growth Trends by Ability Level

An interesting result emerged when comparing the total amount of growth made between the whole group samples and the high ability samples. For the MAP Growth assessment, as the samples became exclusively high ability students, their growth also increased. These results were true for both elementary and middle school cohorts as well as for both mathematics and reading. On average, the high ability samples (95th and 98th percentile samples) grew .27 units more per month in elementary mathematics, and .19 units more per month in middle school mathematics

than their peers who scored at less than the 95th percentile. For reading, the high ability samples also grew more on average, however it was marginal (.02 units and .06 units for elementary and middle school, respectively). These findings directly conflicted with the pattern from the EOG data sets. When comparing the Model 1 estimates, the EOG AG sample grew .04 points less per year in elementary mathematics and .37 points less for middle school mathematics. Again, opposite to the MAP Growth data sets, the EOG AG sample grew substantially less in reading with 1.14 points per year in elementary school and 1.72 points per year in middle school. The trends were similar when comparing the Model 2 and Model 3 estimates except for elementary math where the Model 3 estimates were .08 units higher for the AG sample than the non-AG sample.

Both differing growth trends have been found in previous research. Scammacca et al. (2020) found results that aligned with those of the NC EOG data set portion of this study; additionally they examined the rate of change and found that initially low performing students grew more quickly than initially high performing students. Several researchers have indicated that it is likely that measurement issues could be the reason that high ability student growth could appear lower than students with average ability (McCoach et al., 2013; Ryser & Rambo-Hernandez, 2013; Warne, 2014). McBee et al. (2018) found through a simulation study that students whose academic needs were most closely aligned with classroom instruction (i.e., average ability students in a typical classroom) demonstrated the most growth. Rambo-Hernandez & McCoach (2015) found that although high-ability students grew less during the school year, their growth remained constant during the summer, when average ability student growth staggered. Additionally, Rambo-Hernandez et al. (2021) found that higher ability students initially had lower growth, but over time their growth increased resulting in higher

growth by the end of elementary school. Finally, Warne (2014) found that when using an above-level assessment, high-ability students demonstrated higher growth than the norm group (average ability students). Given these studies in conjunction with the present study, it is possible that high ability students are growing more overall from year-to-year when summer months are taken into consideration, but despite the Tobit corrections, measurement issues with the state achievement tests were still not able to fully capture high-ability student growth. However, if high-ability students are growing more overall due to the summer months, their classroom instruction is not responsible for this growth; a consideration that has serious implications.

Implications For Use of the Tobit Model

Based on the results of this study, the Tobit model is most useful under a certain set of conditions. The Tobit model made the largest corrections when using homogeneous samples of AG students with approximately 15% at the ceiling when lowered to the 95th percentile to account for measurement error. However, if the estimates are too heavily censored, the corrections may not be large enough. The question of the size of corrections made by the Tobit model need to be explored in future research to determine practical significance. Based on these circumstances, the results of this study imply that the use of a multilevel Tobit model for measuring high-ability student growth when the measure used has ceiling effects is best suited for the special population of students identified as Academically or Intellectually Gifted by their local district. However, the results of this study need to be replicated with additional state tests to confirm this implication.

The results of this study also have implications for teacher evaluation; the potential for such was discussed in Chapter Two of this dissertation. Since this study did not explicitly examine value-added estimates, conclusions can be drawn based on the assumption that teachers

of homogeneous groups of high-ability students would benefit from their students demonstrating higher levels of growth. Based on that assumption, the improvement of high-ability student growth using the Tobit models would be useful in improving the outcomes of evaluations for teachers of homogeneous groups of high-ability students. Welsh (2011) noted that when high-ability student growth was inaccurate due to ceiling effects, teachers of these groups of students appeared less effective. Additionally, Resch and Isenberg (2018), Ng and Koretz (2015), and Koedel and Betts (2009) found that in the presence of ceiling effects, value-added estimates for teachers of high-ability students decreased. Therefore, if the growth estimates can be increased via Tobit modeling, there is potential for the value-added estimates to become more accurate.

Ultimately, despite the moderate success of Tobit modeling in this study, the results indicated that for Tobit modeling to be useful, a richer set of covariates that may improve the model's corrections are necessary. Although it would be ideal to use computer-adaptive above-level tests for both measuring and modeling high-ability student growth it would be an expensive overhaul to the testing system. However, if it was possible to use computer adaptive assessments in all states with all students, this would result in more accurate measurement for all students and significantly more accurate measurement for high ability students. Computer adaptive tests have been shown to engage students of all ability levels better, therefore resulting in better performance across the ability spectrum (Martin & Lazendic, 2018). These types of tests have been used with high-ability students and suggested for use for nearly two decades (Adelson & Dickenson, 2016; Assouline et al., 2009; Assouline & Lupkowski-Shoplik, 2012; Lee et al., 2008; Lubinski & Benbow 2006; Rogers, 2002; Swiatek, 2007; Subotnik et al., 2011; Warne, 2012; Warne, 2014). Additionally, Matthews et al. (2012) argued that above-level tests were the best solution for addressing ceiling effects for this population.

Ultimately, although computer-adaptive above-level testing would be an idyllic solution for measuring and modeling high-ability student achievement and growth, it would require an expensive overhaul of the entire testing system across all states that is unlikely to occur. Therefore, based on the results of this study, it is worth exploring the potential use of the Tobit model to solve the aforementioned problems with measurement and modeling of high ability student growth.

Future Research

Despite the successes of the Tobit model under the conditions in the present study, it is important to note that censoring occurs at multiple time points for vertically scaled achievement tests which are censored at each grade level. Therefore, it is important for future research to explore options for accounting for this censoring across time. One R package, *lmmot,* was written by Wright (2014) for the purpose of extending traditional Tobit modeling with only one censor point to account for multiple censor points. Since the publication of Wright and Ziegler (2015) in which the authors used this package, no [English language] articles have cited this package nor the article. This indicates that the extension of Tobit regression to include multiple censor points is new and therefore needs further exploration into its validity and performance. Additionally, that program does not allow for the multilevel structure which is best-suited for educational achievement data. Future research should consider developing a package that allows for both a multilevel structure and multiple censor points.

As previously mentioned, no research has explored the impact of severity of censoring on the performance of the Tobit model, therefore research into this topic should be addressed. Additionally, it would be useful for additional research that focuses solely on the impact of differing percentages of censored observations on Tobit model performance.

Finally, the independent variables used in this study were exclusively demographic variables. Therefore, it would be worth exploring other independent variables that may result in better performance of the Tobit model. One such variable might be formative assessment scores collected by teachers throughout the school year. Logically, performance on other assessments should have a greater impact than demographics on the probability of being censored as well as the correction procedure. Improvement of the performance of the Tobit model (as measured by the size of the model's corrections) should be the most important focus of future research as this line of inquiry would result in the most directly useful information that would allow the Tobit model to be implemented immediately.

Conclusion

Pressures from accountability-driven assessment have resulted in a narrowing of both the curriculum and pedagogy which has driven the rigor of instruction in classroom settings down. In these settings, high-ability students are not in their ZPD and therefore may not learn anything new all year. Despite the lack of rigor in the classroom, these students are still able to reach the measurement ceiling of minimum competency grade-level standardized achievement tests. Therefore, both the achievement and growth of high-ability students as measured by those tests is potentially inaccurate. This study sought to determine if an alternate model, the Tobit model, was able to correct measurement issues for high-ability students with these types of assessments. If these alternate methods of measurement or modeling of high ability student growth are conducted, theoretically teacher pressure associated with testing and meeting growth expectations should also be relaxed. If teachers of high ability students feel less pressure to narrow both their curriculum and pedagogy, the needs of these students are more likely to be met.

The first part of this study imposed artificial ceilings on an above-level computer adaptive assessment that mimicked grade-level ceilings. Tobit modeling was able to come close to replicating the true growth for the high-ability samples of students (those at the 95th and 98th percentiles). Though these models were not perfect, they were able to make up for between 12% and 63% of the censoring that occurred depending on certain conditions such as the percentages of censored observations and the severity of censoring. These results indicated that using Tobit models would be worthwhile with naturally censored data.

The second portion of this study used Tobit modeling on naturally censored accountability-driven standardized state achievement tests. The results of this portion of the study indicated that substantial increases in estimates were made using Tobit modeling when the ceiling was lowered to the 95th percentile to account for other types of measurement error.

Several common trends emerged across both parts of this study. First, the proportion of censored observations needed to be roughly between 5% and 15%. Second, when estimates were more severely censored, the Tobit model was unable to make up for such serious deficits. Additionally, students grew more in mathematics than reading and more in elementary school than middle school; trends which confirmed those of numerous prior studies. Finally, when using the computer adaptive above-level assessment, high-ability students demonstrated higher growth than their average ability peers; however, when using the grade-level state achievement test, AG students demonstrated lower growth than their average ability peers. Given the fact that growth estimates are often used for teacher evaluations, staffing and promotion decisions, and sometimes even monetary bonuses, it is of the utmost importance that growth estimates are accurate for all populations of students.

This study has demonstrated that Tobit modeling has potential under certain circumstances dependent on the proportion of censored observations as well as the severity of censoring. This type of model could be improved by the development of a model that accounts for both the multilevel structure as well as censoring at multiple time points. Additionally, it is important to explore covariates that could be added to the model to improve model performance (i.e., the size of model corrections). Ultimately, the Tobit model is an excellent option that should continue to be explored for its potential to improve the modeling of high-ability student growth.

REFERENCES

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

Amore, M. D., & Murtinu, S. (2019). Tobit models in strategy research: Critical issues and applications. *Global Strategy Journal, 11*(3), 331-355. https://doi.org/10.1002/gsj.1363

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Educational Policy Analysis Archives, 10*(18), 1-74. http://epaa.asu.edu/epaa/v10n18/

Anderman, E. M., Gimbert, B., O'Connell, A. A., & Riegel, L. (2015). Approaches to academic growth assessment. *British Journal of Educational Psychology, 85*(2), 138–153. https://doi.org/10.1111/bjep.12053

APA (n.d.). Ceiling effect. In APA Dictionary of Psychology. Retrieved July, 2021 from https://dictionary.apa.org/ceiling-effect

APA (n.d.). Censored data. In APA Dictionary of Psychology. Retrieved July, 2021 from https://dictionary.apa.org/censored-data

APA (n.d.). Maximum likelihood estimation (MLE). In APA Dictionary of Psychology. Retrieved July, 2021 from https://dictionary.apa.org/mle

Assouline, S. G., Lupkowski-Shoplik, A. (2012). The Talent Search model of gifted identification. *Journal of Psychoeducational Assessment, 30*, 45–59. https://doi.org/10.1177/0734282911433946

Assouline, S., Colangelo, N., Lupkowski-Shoplik, A., Lipscomb, J., Forstadt, L. (2009). *Iowa Acceleration Scale manual (3rd ed.)*. Great Potential Press.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267. http://dx.doi.org/10.3102/0013189X07306523

Austin, P. C., Escobar, M., Kopec, J. A. (2000). The use of the Tobit model for analyzing measures of health status. *Quality of Life Research, 9*, 901-910. https://doi.org/?X?

Barros, M., Galea, M., Leiva, V., & Santos-Neto, M. (2018). Generalized Tobit models: Diagnostics and application in econometrics. *Journal of Applied Statistics, 45*(1), 145-167. https://doi.org/10.1080/02664763.2016.1268572

Bazemore, M., Van Dyk, P., Kramer, L., Yelton, A., & Brown, R. North Carolina Department of Public Instruction (2006). *North Carolina mathematics tests: Technical report.* NCDPI.

Berliner, D. (2011) Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287-302. https://doi.org/10.1080/0305764X.2011.607151

Blazar, D., & Pollard, C. (2017). Does test preparation mean low-quality instruction? *Educational Researcher, 46*(8), 420-433. https://doi.org/10.3102/0013189X17732753

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal, 53*(2), 324-359. https://doi.org/10.3102/0002831216630407

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231–268. http://dx.doi.org/10.3102/00028312042002231

Breen, R. (1996). *Regression models: Censored, sample selected, or truncated data*. Sage.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education, 85*(2), 186–204. http://dx.doi.org/10.1080/01619561003685346

Callahan, C. M. (2009). Making the grade or achieving the goal: Evaluating learner and program outcomes in gifted education. In F. A. Karnes & S. M. Bean (Eds.), *Method and materials for teaching the gifted* (3rd ed., pp. 221–258). Prufrock Press.

Cao, T. H., Jung, J. Y., & Lee, J. (2017). Assessment in gifted education: A review of the literature from 2005 to 2016. *Journal of Advanced Academics, 28*(3), 163–203. https://doi.org/10.1177/1932202X17714572

Castellano, K. E., Ho, A. D. (2013). *A practitioner's guide to growth models.* Washington, DC: Council of Chief State School Officers.

Cowles, M. K., Carlin, B. P., Connett, J. E. (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association, 91*(433), 86-98. https://doi.org/10.1080/01621459.1996.10476666

Crocco, M. S., & Costigan, A. T. (2007). The narrowing of curriculum and pedagogy in the age of accountability: Urban educators speak out. *Urban Education, 42*(6), 512-535. https://doi.org/10.1177%2F0042085907304964

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Holt, Rinehart, and Winston.

Davis, G. A., Siegle, D. B., & Rimm, S. B. (2018). *Education of the gifted and talented* (7th ed.). Pearson.

Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education, 80*(4), 285–313. http://dx.doi.org/10.1177/003804070708000401

Egan, K. L., & Davidson, A. H. (2017, Nov. 14). *Alignment of the NWEA MAP Growth & MAP Growth K–2 to the Common Core State Standards: English language arts & mathematics.* EdMetric

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015).

Farkas, S., & Duffett, A. (2008). *High achieving students in the era of NCLB: Results from a national teacher survey*. Washington, DC: Thomas E. Fordham Institute.

Farvis, J., & Hay, S. (2020). Undermining teaching: How education consultants view the impact of high-stakes test preparation on teaching. *Policy Futures in Education, 18*(8), 1058-1074. https://doi.org/10.1177%2F1478210320919541

Fuller, S. C., & Ladd, H. F. (2013). School-based accountability and the distribution of teacher quality across grades in elementary school. *Education Finance and Policy, 8*(4), 528-559. https://www.jstor.org/stable/10.2307/educfinapoli.8.4.528

Gallagher, J. J. (2004). No child left behind and gifted education. *Roeper Review, 26*(3), 121-123. http://dx.doi.org/10.1080/02783190409554255

Gentry, M. (2006). No child left behind: Neglecting excellence. *Roeper Review, 29*(1), 24-27. https://doi.org/10.1080/02783190609554380

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Washington, DC: Council of Chief State School Officers.https://files.eric.ed.gov/fulltext/ED542761.pdf

Guarino, C. M., Reckase, M. D., & Woolridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy, 10*(1), 117-156. https://www.jstor.org/stable/10.2307/educfinapoli.10.1.117

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Kluwer Nijhoff Publishing.

Hawley, L. R., Bovaird, J. A., & Wu, C. R. (2017). Stability of teacher value-added rankings across measurement model and scaling conditions. *Applied Measurement in Education, 30*(3), 196-212. https://doi.org/10.1080/08957347.2017.1316273

Heckman, J. J. (1974). Shadow prices, market wages and labor supply. *Econometrica, 42* (4), 679-694. https://doi.org/10.2307/1913937

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection bias, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement, 5* (4), 475-492. https://www.nber.org/system/files/chapters/c10491/c10491.pdf

Hendrickson, A. (2007). *An NCME instructional module on multistage testing.* Instructional Topics in Educational Measurement. NCME.

Ho, A. D., (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37* (6), 351-360. https://doi.org/10.3102/0013189X08323842

Hujar, J. (2021, November). *The impact of graduate-level training in gifted education on perceptions of assessment.* [Poster]. National Association for Gifted Children Annual Conference. Denver, Co.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004)

Jaeger, R. M., Tittle, C. K. (1980). *Minimum competency testing: Motives, models, measures and consequences.* McCutchan.

Jerald, C.D. (2006). *The hidden costs of curriculum narrowing*. Washington, DC: Learning Point Associates Issue Brief. The Center for Comprehensive School Reform and Improvement. http://www.centerforcsri.org/files/CenterIssueBriefAug06.pdf

Kaplan, S. N. (2004). Where we stand determines the answers to the question: Can the no child left behind legislation be beneficial to gifted students? *Roeper Review, 26*(3),124-125. https://doi.org/10.1080/02783190409554256

Koedel, C., & Betts, J. (2009). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *National Bureau of Economics Research,* Working Paper 14778. http://www.nber.org/papers/w14778

Koretz, D. M. (2008). *Measuring* up. Harvard University Press.

Koretz, D., Hamilton, L. (2006). Testing for accountability in K–12. In Brennan, R. L. (Ed.), *Educational measurement* (5th ed., pp. 531–578). American Council on Education/Praeger.

Lee, S., Matthews, M. S., & Olzewski-Kubilius, P. (2008). A national picture of talent search and talent search educational programs. *Gifted Child Quarterly, 52* (1), 55-69. https://doi.org/10.1177/0016986207311152

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Sage.

Lord, F. M. (1974). *Practical methods for redesigning a homogeneous test, also for designing a multilevel test*. Educational Testing Service RB-74–30.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum.

Loyd, B. H. (1984). *Efficiency and precision in two-stage adaptive testing*. Eastern ERA.

Lubinski, D., Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science, 1*, 316–345. https://doi.org/10.1111/j.1745-6916.2006.00019.x

Maker, C. J., & Schiever, S. W., (2005). *Teaching models in education of the gifted* (3rd ed.). PRO-ED.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–74.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27-45. https://doi.org/10.1037/edu0000205

McBee, M. T. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly, 54*(4), 314–320. https://doi.org/10.1177/0016986210379095

McBee, M. T., & Makel, M. C. (2019). The quantitative implications of definitions of giftedness. *AERA Open, 5*(1), 1-13. https://doi.org/10.1177%2F2332858419831007

McCabe, E., Gross, D. P., Bulut, O. (2018). Procedures to develop a computerized adaptive test to assess patient-reported physical functioning. *Quality of Life Research, 27*, 2393-2402. https://doi.org/10.1007/s11136-018-1898-0

McCoach, D. B., Madura, J. P., Rambo-Hernandez, K. E., O;Connell, A. A., & Welsh, M. E. (2013). Longitudinal data analysis. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (199-230). Sense Publishers. http://ebookcentral.proquest.com/lib/uncc-ebooks/detail.action?docID=3034937

McCoach, D. B., Peters, S. J., Richardson, J. (2021, April 10). *Waiting to learn? An exploratory examination of achievement growth (and achievement gaps) by initial proficiency.* [Paper presentation]. American Educational Research Association Annual Meeting, Virtual.

McCoach, D. B., Rambo, K. E., Welsh, M. (2013). Assessing the growth of gifted students. *Gifted Child Quarterly, 57*(1), 56–67. https://doi.org/10.1177/0016986212463873

McDonald, J. F., & Moffit, R. A. (1980). The uses of Tobit analysis. *The Review of Economics and Statistics, 62*(2), 318-321. https://www.jstor.org/stable/1924766

McNeill, L. (2002). *Contradictions of school reform: Educational costs of educational testing*. Routledge.

McNeish, D., & Matta , T. (2018). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior Research Methods, 50*, 1398–1414. https://doi.org/10.3758/s13428-017-0976-5

Moon, T. R., Brighton, C. M., & Callahan, C. M. (2003). State standardized testing programs: Friend or foe of gifted education? *Roeper Review, 25*(2), 49-60. https://doi.org/10.1080/02783190309554199

Myers, C. (2012). The centralizing role of terminology: A consideration of achievement gap, NCLB, and school turnaround. *Peabody Journal of Education, 87*, (4), 468-484. https://doi.org/10.1080/0161956X.2012.705149

National Association for Gifted Children. (2019). *Position paper: A definition of giftedness that guides best practice*. NAGC. https://www.nagc.org/sites/default/files/Position%20Statement/Definition%20of%20Giftedness%20%282019%29.pdf

Ng, H. L., & Koretz, D. (2015). Sensitivity of school-performance ratings to scaling decisions. *Applied Measurement in Education, 28*(4), 330-349. https://doi.org/10.1080/08957347.2015.1062764

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

North Carolina Department of Public Instruction (2006). *The North Carolina mathematics tests: Technical report*. NCDPI.

North Carolina Department of Public Instruction (2009). *The North Carolina reading comprehension tests: Technical report.* NCDPI.

North Carolina Department of Public Instruction. (2017). *North Carolina ESSA state plan*. NCDPI. https://www.dpi.nc.gov/media/8459/download

NWEA. (2016). *Linking the North Carolina EOG assessments to NWEA MAP Growth tests*. NWEA.

NWEA. (2019). *MAP® Growth™ technical report.* NWEA.

Peters, S. J., Rambo-Hernandez, K. E., Makel, M. C., Matthews, M. S., Plucker, J. A. (2016). Should millions of students take a gap year? Large numbers of students start the school year above grade level. *Gifted Child Quarterly, 61*(3), 229–238. https://doi.org/10.1177/0016986217701834

Peters, S. J., Rambo-Hernandez, K. E., Makel, M. C., Matthews, M. S., Plucker, J. A. (2019). Effect of local norms on racial and ethnic representation in gifted education. *AERA Open, 5*(2), 1-18. https://doi.org/10.1177/2332858419848446

Plucker, J. A., & Callahan, C. (2014). *Critical issues and practices in gifted education: What the research says* (2nd ed.). Prufrock Press.

Plucker, J. A., & Callahan, C. (2014). *Research on giftedness and gifted education: Status of the field and considerations for the future.* Exceptional Children, 80(4), 390-406. https://doi.org/10.1177%2F0014402914527244

Popham, W. J. (2001). Teaching to the test? *Educational Leadership, 58*(6), 16–21.

Ragosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726–748. https://doi.org/10.1037/0033-2909.92.3.726

Rambo-Hernandez, K. E., & McCoach, D. B. (2015). High-achieving and average students' reading growth: Contrasting school and summer trajectories. *The Journal of Educational Research, 108*(2), 112-129. https://doi.org/10.1080/00220671.2013.850398

Rambo-Hernandez, K. E., & Warne, R. T. (2015). Measuring the outliers: An introduction to out-of-level testing with high-achieving students. *Teaching Exceptional Children*, *47*(4), 199-207. https://doi.org/10.1177%2F0040059915569359

Rambo-Hernandez, K. E., Makel, M., Peters, S. J., & Worley, C. (2021, April 10). *Differential return on investment: Academic growth based on initial performance.* [Paper presentation] American Educational Research Association Annual Meeting, Virtual.

Rambo-Hernandez, K. E., Peters, S. J., Plucker, J. A., Makel, M. C., Pederson, B. (2021, April 10). *How academically diverse is the "grade-level" classroom?* [Paper presentation]. American Educational Research Association Annual Meeting, Virtual.

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

Renzulli, K. S. (1986). *Systems and models for developing programs for the gifted and talented*. Creative Learning Press.

Resch, A., & Isenberg, E. (2018). How do test scores at the ceiling affect value-added estimates? *Statistics and Public Policy, 5*(1), 1-6. https://doi.org/10.1080/2330443X.2018.1460226

Resnick, L. B. (2010). Nested learning systems for the thinking curriculum. *Educational Researcher, 39*(3), 183-197. https://doi.org/10.3102%2F0013189X10364671

Robins, J. H., Jolly, J. L., Karnes, F. A., & Bean, S. M. (2020). *Methods & materials for teaching the gifted* (5th ed.). Prufrock Press.

Rogers, K. B. (2002). *Re-forming gifted education: How parents and teachers can match the program to the child.* Great Potential Press.

Rooney, E. (2015). "I'm just going through the motions": High-stakes accountability and teachers' access to intrinsic rewards. *American Journal of Education, 121*(4), 475-500. https://www.jstor.org/stable/10.1086/681923

Ryser, G. R., Rambo-Hernandez, K. E. (2013). Using growth models to measure school performance: Implications for gifted learners. *Gifted Child Today, 37*(1), 17-23. https://doi.org/10.1177/1076217513509617

SAGE (n.d.). Value added models (VAM). In SAGE Encyclopedia of Educational Research, Measurement, and Evaluation . Retrieved July, 2021 from https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i21957.xml

SAGE (n.d.). Vertical scaling. In SAGE Encyclopedia of Educational Research, Measurement, and Evaluation . Retrieved July, 2021 from https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i22066.xml

SAS. (2021). *SAS EVAAS for K-12 statistical models*. SAS.

Shepard, L. (2002–2003, Winter). The hazards of high stakes testing. *Issues in Science and Technology*, pp. 53–58.

Sigelman, L., & Zeng, L. (2000). Analyzing censored and sample-selected data with Tobit and Heckit models. *Political Analysis, 8*(2), 167-182. https://www.jstor.org/stable/25791605

Smithson, J. L. (2015). *A report to the North Carolina department of public instruction on the alignment characteristics of state assessment instruments covering grades 3-8, and high school in mathematics, reading and science.* Wisconsin Center for Educational Research. https://www.dpi.nc.gov/media/9629/open

Springer Link (n.d.). Multilevel Tobit regression. In Springer Link Encyclopedia of Quality of Life and Well-Being Research. https://link.springer.com/referenceworkentry/10.1007%2F978-94-007-0753-5_3025

Swiatek, M. A., Lupkowski-Shoplik, A. (2005). An evaluation of the elementary student Talent Search by families and schools. *Gifted Child Quarterly, 49*, 247–259. https://doi.org/10.1177/001698620504900306

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment and school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–36. https://doi.org/10.3102/10769986029001011

The University of North Carolina at Charlotte. (n.d.). *Graduate certificate in academically or intellectually gifted.* Distance Education UNC Charlotte. https://distanceed.charlotte.edu/programs/academically-or-intellectually-gifted

Thum., Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth*. NWEA.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*(1), 24-36. http://links.jstor.org/sici?sici=0012-9682%28195801%2926%3A1%3C24%3AEORFLD%3E2.0.CO%3B2-R

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201.

Walker, C. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment, 29*, 364-376

Wang, L. & Zhang, Z. (2011). Estimating and testing mediation effects with censored data. *Structural Equation Modeling, 18*(1), 18-34. https://doi.org/10.1080/10705511.2011.534324

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research, 43*(3), 476-496. https://doi.org/10.1080/00273170802285941

Wang, S., Jiao, H., & Zhang, Z. (2013). Validation of longitudinal achievement constructs of vertically scaled computerized adaptive tests: A multiple-indicator, latent-growth modelling approach. *International Journal of Quantitative Research in Education, 1*(4), 383–407.

Wang, S., McCall, M., Jiao, H., & Harris, G. (2013). Construct validity and measurement invariance of computerized adaptive testing: Application to Measures of Academic

Progress (MAP) using confirmatory factor analysis. *Journal of Educational and Developmental Psychology, 3*(1), 88–100

Wang, W., & Griswold, M. E. (2017). Natural interpretations in Tobit regression models using marginal estimation methods. *Statistical Methods in Medical Research, 26*(6), 2622-2632. https://doi.org/10.1177%2F0962280215602716

Warne, R. T. (2012). Historic views and current understandings of above-level testing. *Roeper Review, 34*(3), 183-193. https://doi.org/10.1080/02783193.2012.686425

Warne, R. T. (2014). Using Above-Level Testing to Track Growth in Academic Achievement in Gifted Students. *Gifted Child Quarterly, 58*(1), 3–23. https://doi.org/10.1177/0016986213513793

Weiss, D. J. (1974). *Strategies of adaptive ability measurement.* Psychometric Methods Program, Research Report 74–5, Department of Psychology, University of Minnesota, Minneapolis.

Welsh, M. E., Eastwood, M. & D'Agostino, J. V. (2014). Conceptualizing teaching to the test under standards-based reform. *Applied Measurement in Education, 27*(2), 98-114. https://doi.org/10.1080/08957347.2014.880439

Wright, M. N. (2014). lmmot: Multiple Ordinal Tobit (MOT) model. R package version 0.1.2. http://CRAN.R-project.org/package=lmmot

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly, 10*(2), 196-218. https://doi.org/10.1080/15434303.2012.721423

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology*(Vol. 5, pp. 269–304). Emerald Group

Zwick, R., Thayer, D.T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement, 36*(1),1-28.

APPENDIX A: Tobit Correction

After categorizing observations as censored and uncensored, a linear function of the independent variables called an index function (I_t) is created, where X is a vector of independent variables, *t* is the individual, β is the regression coefficient, and σ is the standard error.

$$I_t = X_t\alpha = X_t\left(\frac{\beta}{\sigma}\right)$$

The Tobit equation is estimated indirectly using MLE of the following log-likelihood equation where F is the normal cumulative distribution function and *f* is the normal probability density function.

$$L = \sum_{t=1}^{N} log[1 - F(\sigma\, Y_t - I_t)] + \sum_{t=N+1}^{T} log f(\sigma Y_t - I_t)$$

F and *f* are calculated at I_t and then the following equation is used to calculate the conditional expectation of Y_t.

$$E(Y_t|I_t) = \sigma I_t F(I_t) + \sigma f(I_t)$$

The estimates for the uncensored dependent variable are calculated using the following equation:

$$E(Y_t|I_t, Y_t < \tau) = \sigma I_t + \frac{\sigma f(I_t)}{F(I_{t)}}$$

Lastly, the predicted value of the dependent variable is calculated using this final equation:

$$\hat{Y}_t = \hat{\sigma}\hat{I}_t F(\hat{I}_t) + \hat{\sigma} f(\hat{I}_t)$$

(Adapted from Ekstrand & Carpenter, 1998)