2016

Missing Data in the Context of Student Growth

Katherine Wright
Loyola University Chicago

LOYOLA UNIVERSITY CHICAGO

MISSING DATA IN THE CONTEXT OF STUDENT GROWTH

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

PROGRAM IN RESEARCH METHODOLOGY

BY

KATHERINE M. WRIGHT

CHICAGO, IL

DECEMBER 2016

ACKNOWLEDGEMENTS

It is with tremendous gratitude that I thank several people, without whom this dissertation would not be possible.

I want to thank my advisor, Dr. Terri Pigott, for her methodological expertise, and for seeing this project through despite an increasingly busy schedule. Her research on missing data and meta-analysis is inspirational. I owe a great deal to the passionate and exceptionally supportive professors in the Research Methodology program at Loyola University Chicago, particularly Drs. Wu and Kallemeyn.

I would also like to thank Dr. John Gatta for serving on my committee, letting me TA two of his biostatistics courses, and being a mentor I truly respect. I also need to thank my friends and colleagues in the Department of Family & Community Medicine at the Northwestern University Feinberg School of Medicine for their incredible support and guidance. I feel lucky to have had the opportunity to work with such wonderful people - Tim Doyle, Lauren Anderson, Arona Gur, Dr. Elizabeth Ryan, and Dr. Deborah Clements, to name a few.

Special thanks to Heather Pease for the camaraderie and encouragement along the way; this journey would not have been the same without you.

To my parents, for their untiring support in all aspects of life, but especially for impressing upon me the importance of education. They have sacrificed so much, and I am forever indebted. To my brother Thomas, for inspiring me to be a sister he could be

proud of. To Mark, for his support in good times and in bad over the last decade; thank you for your unwavering belief in my ability to see this process through to the end.

Due to the kindness of others, this work was possible. I am eternally grateful to you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

One property of student growth data that is often overlooked despite widespread prevalence is incomplete or missing observations. As students migrate in and out of school districts, opt out of standardized testing, or are absent on test days, there are many reasons student records are fractured. Missing data in student growth models can bias model estimates and growth inferences. This study presents empirical explorations of how well missing data methodologies recover attributes of would-be complete student data used for teacher evaluation. Missing data methods are compared in the context of a Student Growth Percentiles (SGP) model used by several school systems for accountability purposes. Using a real longitudinal dataset, this study evaluates the sensitivity of growth estimates to missing data and compares the following missing data methods: listwise deletion, likelihood-based imputation using an expectation-maximization algorithm, multiple imputation using a Markov Chain Monte Carlo method, multiple imputation using a predictive mean matching method, and inverse probability weighting. Methodological and practical consequences of missing data are discussed.

CHAPTER ONE

INTRODUCTION

Now more than ever, policymakers and researchers alike are interested in measuring a teacher's contribution to student learning. This attention stems from the basic notion that teacher quality drives student achievement. Historical frameworks for teacher evaluation resulted with a majority of teachers receiving the top proficiency rating; as the secretary of education highlights, "99% of our teachers are above average (Gabriel, 2010)." Despite consistent educator ratings, student experiences vary considerably by location, demographics, and socioeconomic status, among other factors (Aud et al., 2011). When every educator receives the same rating it becomes impossible to make decisions based on evaluations, making the evaluation process a formality instead of a tool for continuous improvement. Recognizing differences among districts, schools, and teachers is essential in making informed decisions about best pedagogical practices and adequate student progress. As school systems look for ways to better identify effective teachers, conversations around accountability are increasingly centered on standardized test scores and the inferences that can be made from them.

Measurement approaches broadly categorized as "value-added" growth models (VAMs) attempt to quantify teacher effectiveness while accounting for baseline characteristics like prior achievement through advanced statistical techniques. Economists first used value-added models to explore the effect of class size and other

controllable factors on student achievement (Koedel, Mihaly, & Rockoff, 2015). VAMs were used to identify the most impactful way to spend limited resources by quantifying school systems' return on investment. VAM methodology has since extended to teacher evaluation. Often VAMs are meant to partial out a teacher's contribution towards a student's growth, with the difference between a student's actual and predicted score representing their teacher's value added contribution.

Despite the prominence of VAMs, there are competing views on appropriate value-added measurement and inference (Amrein-Beardsley, 2008). Methodologists continually highlight model limitations and refine statistical techniques. Some worry unmeasured variables may result in biased models and unfair evaluation systems, particularly for teachers of disadvantaged groups if not accounted for by the model. Others worry setting differential expectations can sustain or even contribute to the achievement gap (Ballou, Sanders, & Wright, 2004). However, the conceptual appeal of VAMs perpetuates their use across the nation.

Decisions regarding VAMs are complicated and multifactorial. Given the wide variety of methods and uses for value-added modeling, it is difficult to arrive at a set of best practices for specifying a model or evaluation system. Addressing this issue, in November 2015 the American Educational Research Association (AERA) released a statement outlining 8 technical requirements to guide use of VAM in educator evaluations ("AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs," 2015). In response, the Brookings Institute suggests the educational community "must view the value of any

particular performance measure in the context of all other measures, not relative to a nirvana that does not exist (Hansen, 2015)."

In 2010, the LA Times released school and educator effectiveness rankings derived from a district-wide VAM ("Los Angeles Teacher Ratings," 2010). This sparked controversy as schools and teachers expressed concern over making these ratings public given the methodology used to rate schools and teachers is an imperfect science (Briggs & Domingue, 2011). One concern centers on the variation in VAM estimates attributable solely to model specification. Re-analyzing the LA student achievement data, the National Education Policy Center reported using the same VAM but controlling for additional factors resulted in a .92 correlation between the two model estimates (Goldhaber, Walch, & Gabele, 2014). However, applying this change to the evaluation framework resulted in inconsistent effectiveness ratings for 40% of math teachers between the two VAMs. A relatively small statistical adjustment could translate to drastically different evaluation inferences.

Due to the high-stakes nature of evaluation, inference decisions should be grounded in sound methodology. The accuracy of any statistical model relies on the extent to which certain assumptions are met, and the same is true of VAMs. The complex school environment, students' non-random assignment to classrooms, and immeasurable variables that influence student learning all make it important to closely investigate the statistical properties of each model to accurately interpret its results. One property of assessment data that is often overlooked despite widespread prevalence is incomplete or missing student observations.

Most VAMs are designed for complete datasets. Consequently, analyzing student achievement data prone to missing observations may impact model findings and subsequent data-driven decisions. As students migrate in and out of school systems, leave high school, or are absent from testing, there are many reasons student records are fractured. If not properly accounted for, incomplete student data may be an invisible covariate affecting evaluation inferences in student growth models. Further research is necessary to ensure student growth models mitigate bias due to missing data.

Though there is still healthy debate regarding VAM methodology and best practices, The American Statistical Association points out "under some conditions, VAM scores and rankings can change substantially when a different model or test is used, and a thorough analysis should be undertaken to evaluate the sensitivity of estimates to different models (American Statistical Association, 2014)." This dissertation will evaluate the sensitivity of VAM estimates to missing data and methods used to account for missing data.

Research Questions

The purpose of this dissertation is to present empirical explorations of how well missing data methodologies recover attributes of would-be complete student data used for teacher evaluation. In studying this topic, both methodological and practical consequences of missing data are of interest. Using a longitudinal dataset of student records, this research will address the following:

1. How sensitive are growth estimates to missing data?
2. Does the choice of missing data methodology result in different growth inferences when used in an educator evaluation framework?

This dissertation will not advocate for a single missing data handling technique, nor will it present evidence demonstrating the superiority of a particular method for universal use in every VAM application. Further, it will not quell or ignite the larger debate surrounding VAM methodology. Missing data procedures and value-added growth modeling procedures in general are inherently neutral. Users must subjectively derive meaning from objective statistical output, as models alone cannot produce a central argument favoring one decision or another. Rather, VAMs produce evidence, and methodologists and school systems are left to evaluate the quality of evidence before making inference decisions. This study will contribute to the growing body of evidence around value-added growth models with respect to missing data.

CHAPTER TWO

LITERATURE REVIEW

Teachers are evaluated for tenure, promotion, compensation, contract renewal, corrective-action and dismissal. Historically principal observations were the most popular method of evaluating teacher effectiveness. Like any evaluation system, observational methods require adequate data to make evidence-based decisions. Often teachers aren't observed regularly and without this data the evaluation system is unproductive. Reviewing the frequency of teacher observations in Boston, only 53% of teachers received an evaluation over a two year period (National Council on Teacher Quality, 2010). This figure speaks only to the presence or absence of an evaluation, not to its thoroughness. Observations can be highly subjective and often produce very little variation. Analyzing teacher evaluations in four states, the New Teacher Project found 99% of educators received satisfactory ratings for evaluations based solely on classroom observation (Weisberg et al., 2009). This is problematic as both excellence and ineffectiveness are indistinguishable. An evaluation framework with little variation misses the opportunity for feedback as most teachers receive the same evaluation despite different pedagogy. This system devalues the evaluation process, failing students and teachers.

In search of a more objective evaluation approach, many states incorporate student achievement data to supplement other evaluation components. No Child Left

Behind (NCLB) set requirements for assessment and accountability, requiring students in grades 3-8 to take annual standardized tests. Thresholds must be met for schools to demonstrate adequate yearly progress (AYP) towards the goal of demonstrating proficiency for all students in reading and math. Failure to meet adequate progress resulted in serious consequences and mandatory corrective action. Though NCLB was recently replaced by the Every Student Succeeds Act (ESSA), its legacies are carried forward in the current educational landscape (United States Department of Education, 2015). To this end, school systems are under increasing pressure to demonstrate their "value added" for student achievement, conceptualized primarily by gains in standardized tests.

Unconditional achievement scores, or status metrics, provide valuable information regarding a student's absolute standing defined by an assessment rubric. Status-based accountability systems evaluate teachers and schools based on the percentage of students that meet minimum scores for proficiency status on state-mandated exams with the goal of eventually reaching 100%. However, various factors outside a teacher's control can influence student achievement (Hoff, 2003; Jeynes, 2007; Lee & Burkam, 2002). Because socioeconomic status and other environmental factors play a role in learning, some argue a fair evaluation framework must take background information and prior test scores into account for an evaluation system to be equitable. Without these considerations, teachers in the most high-risk classrooms would be unfairly penalized. Lower-achieving students of highly effective teachers may go unrecognized if they fail to meet proficiency despite making substantial progress. Additionally, it is more difficult to

bring a child up to proficiency than it is to maintain proficiency, placing a heavier burden on schools of disadvantaged student populations (Neal & Schanzenbach, 2010).

Furthermore, there is mixed evidence that status-based accountability systems incentivize schools to target students on the cusp of meeting proficiency standards at the expense of their counterparts on the fringe (Ballou & Springer, 2008). Under proficiency-based systems, schools generate greater return on investment with programs aimed at modest gains for students in the middle with potential to cross the proficiency threshold, as opposed to programs for low- or high-scoring students who aren't likely to affect the overall proficiency rating. For these reasons, school systems turned to growth metrics to supplement status measures in demonstrating AYP. Most evaluation frameworks include multiple components in addition to student growth, acknowledging growth models aren't designed to measure every contribution a teacher makes toward student learning.

Defining Growth

Achievement scores are meant to quantify a student's attainment at a single point in time. Conditional achievement scores, or growth metrics, are meant to provide information about progress over time. Unlike status, however, the concept of growth is less concrete. Growth can be challenging to define and even more challenging to measure. Given the abundance of definitions and measurement approaches in use, growth model terminology is often ambiguous, contributing to the confusion and controversy surrounding VAM implementation. Since there is no common definition for the term "value-added model," for the purpose of this paper VAMs represent a broad category of statistical models used to evaluate growth. Examples include Student Growth

Percentiles models developed by Damian Betebenner and the National Center for the Improvement of Educational Assessment (NCIEA), Multivariate Response Models developed by SAS, and value-added models developed by the Value-Added Research Center (VARC).

Because growth metrics serve a variety of purposes, it is first necessary to settle on a desired end goal. Purposes for modeling can be descriptive or inferential in nature, but should be explicit in either case (Seltzer, Frank, & Bryk, 1994). Some are designed to project future performance (e.g. projection to proficiency 3 to 5 years out) whereas others aim to quantify past student growth. Some frameworks measure growth relative to a criterion and others measure a student's standing relative to their peer group.

Criterion-referenced growth measures anchor progress to a specific content area or domain. The underlying construct of any assessment is achievement, which can be conceptualized as a latent variable (e.g. reading proficiency) indirectly observed through test items and summarized by assessment scores (Cyr & Davies, 2005). To measure growth over time, vertically linked assessments are a series of tests designed to quantify a student's achievement across grades (Lissitz & Huynh, 2003). Raw scores are standardized and equated to a common scale. Though tests administered to different grades cover different content (e.g. mathematics concepts), they measure the underlying concept of mathematics proficiency. Cross-scaling techniques allow for continuous tracking of student achievement as they advance to different grades. Since it is expected students will increase in mathematics proficiency each year, we would expect mathematics scores on vertically linked assessments to increase as well. Because vertical

scales are interval measurements, a student's prior grade score could be subtracted from their current score to calculate their gain from one year to the next.

Though vertical scaling provides a framework to interpret student scores, many parents, teachers, and administrators are left wondering what gain is considered normal or adequate. Test publishers provide recommendations and experts can answer these questions qualitatively, but there is no definite answer for stakeholders. Instead, it can be helpful to frame growth relative to student peers.

Norm-referenced growth measures provide information about an individual's achievement compared to students of a similar test history or background. Betebenner analogizes achievement growth to pediatric weight or height growth to answer questions about what constitutes typical or average growth (D. W. Betebenner, 2008). A 2-pound weight gain may not mean as much to parents without knowing this places their child at the 99th percentile for weight. Similarly, what does a 5-point scale score achievement gain represent in terms of content mastery? Knowing a student's 5-point increase places them at the 95th percentile and translates to performance equal or better than 95% of their peer group provides stakeholders a reference point.

Both normative and criterion-based measurement frameworks should be thoughtfully explored to avoid dangerous misinterpretations. For example, moving from a score of 15 to 20 may be more difficult than moving from 10 to 15 on the same assessment despite equal 5-point gains in both scenarios. The magnitude of achievement growth may be imprecisely captured by an equal interval scale score gain. Since the underlying construct is latent this idea is difficult to confirm, though methodologists can establish an unequal likelihood to achieve equal gains from different starting points along

the baseline distribution. Gain score calculations assume constant variance and are often negatively correlated with a student's initial standing. This phenomenon occurs with widely administered assessments like ACT® as students with a lower baseline score tend to show greater gains than students starting in the middle or high end of the baseline assessment (Andrews & Ziomek, 1998). Similarly, in a normative framework the 50th percentile for the lower end of the distribution may represent a 3-point scale score increase whereas the 50th percentile for the middle of the distribution may represent a 1-point scale score increase. Despite these nuances, when statistical underpinnings of growth models are clearly defined and understood, the information yielded can be a valuable tool for identifying effective programs and pedagogy.

Growth for Accountability

Regardless of definition, growth is increasingly relevant to school and educator evaluations to paint a more complete picture of student progress. This movement gained momentum when the Obama administration incentivized states to link student achievement outcomes to teacher evaluations under the Race to the Top (RttT) initiative (McGuinn, 2011). The RttT announcement coincided with the financial crisis, further incentivizing schools with limited or diminishing resources to compete for government funding. Following NCLB, RttT, Teacher Incentive Fund (TIF) grants, and other state mandates, student achievement data now plays a more prominent role in teacher evaluation than ever before (Linn, Baker, & Betebenner, 2002). However, states were given the flexibility to decide how student achievement data should factor into larger evaluation systems, resulting in a plethora of approaches. New ESSA legislation further

emphasizes state and local responsibility for accountability measures (United States Department of Education, 2015).

Proponents of VAMs believe complex statistical modeling brings objectivity to the evaluation process instead of relying solely on subjective observational ratings. Just as principal observations may be biased in identifying effective teachers, VAMs must have valid, reliable student achievement data in order for the model to accurately quantify growth and facilitate evaluation inferences. Given the high-stakes decisions made from growth data, critics of VAMs point out several methodological limitations about the value-added modeling process. As is true of any statistical model, growth models are only as good as their predictors. Many standardized assessments built to demonstrate school accountability under status models may not be useful assessments to measure student growth (e.g., Steering Committee of the Delaware Statewide Academic Growth Assessment Pilot, 2007). Critics of growth models often view computationally intense analyses as lacking transparency (Ladd & Lauen, 2010). Most relevant to this paper, statisticians point out that growth models, like most statistical procedures, were designed to analyze complete data.

Compared to status metrics alone, VAMs can provide a more complete understanding of a teacher's impact by measuring student progress in comparison to a student's predicted trajectory. An ideal model accounts for the relationships between student characteristics and growth, so that growth scores are not correlated with demographics or initial achievement levels. However, prior research demonstrates VAMs are sometimes correlated with status measures (McCaffrey & Castellano). Nor are growth and status competing frameworks. Raising the minimum proficiency for all

students will remain the ultimate goal of any school system that implements a growth model.

Approaches to Measuring Growth

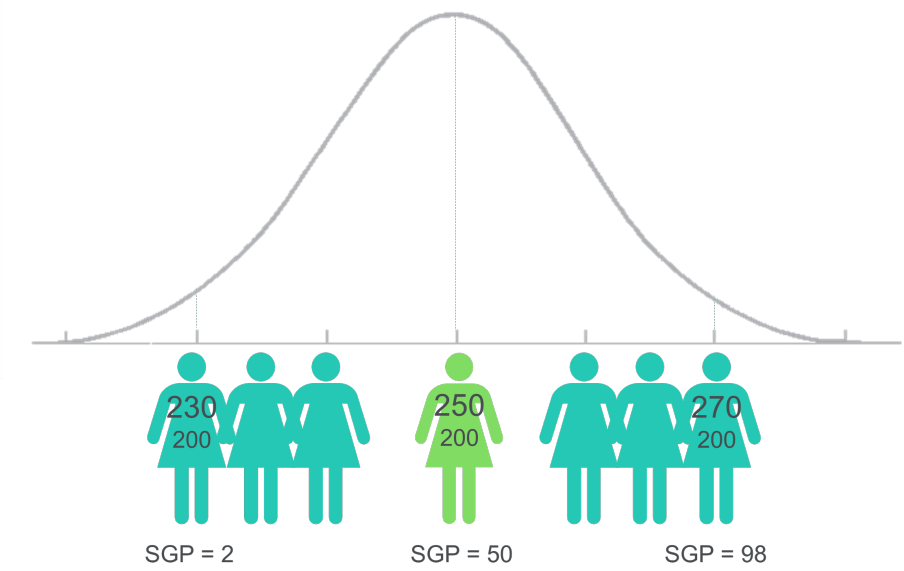
If a school system chooses to include growth data in their accountability system, there are a number of models available to measure progress. Rather than devote resources to develop new models, many implement or modify existing VAMs to measure growth. Ranging from conceptually simple fixed-effects models to more complex longitudinal mixed models, the statistical underpinnings of each model varies. This leaves many considerations for specifying a model. Some include demographic information while others deliberately leave this information out as not to set differential expectations based on ethnicity or other student attributes. School systems must also make decisions regarding how many years of historical data to include in the model and the length of time for measuring growth (e.g. spring to spring models vs. fall to spring). VAMs also differ in how they establish teacher effects (e.g. aggregating student gains or including a teacher term in model). Then they must decide how to incorporate growth data into an evaluation framework by determining acceptable growth thresholds and then weighting the growth component with other evaluation data so they can derive meaning from the information gained through value-added modeling. The focus of this study is the choice of missing data method to account for incomplete student records in value-added models.

Student Growth Percentiles

The Student Growth Percentiles (SGP) model developed by Betebenner (D. W. Betebenner, 2011) was selected as a focus of this review since over 30 states have chosen

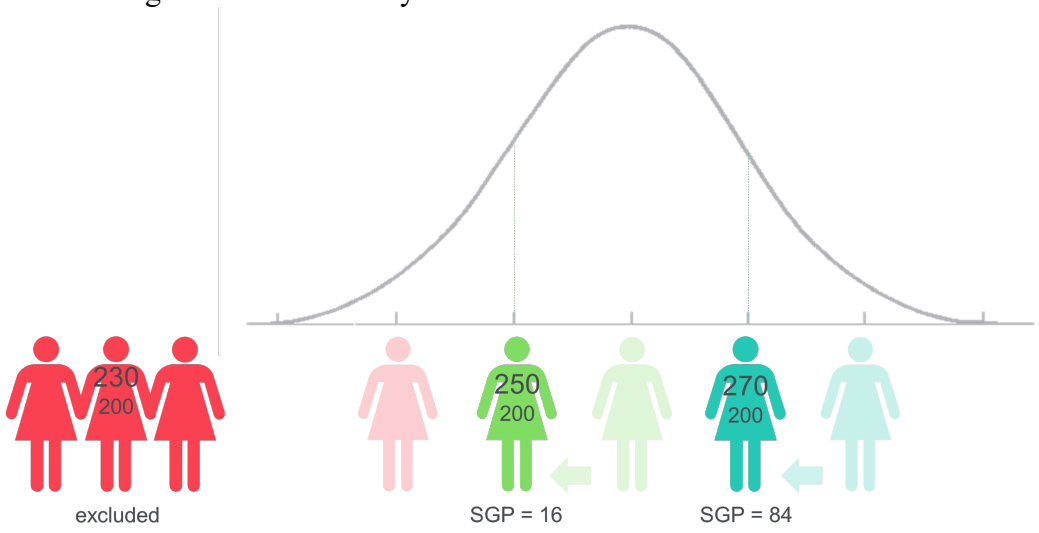
to adopt it in some capacity. This model is normative in nature and produces student percentile ranks as its growth metric. As illustrated in Figure 1, students' current year performance is evaluated by their relative performance to peers with similar assessment histories (in this example, a prior year score of 200).

Figure 1: An Illustration of the Student Growth Percentiles Framework



In this example, a score of 250 translates to an SGP of 50, or median performance among similar students. A normative growth framework presents unique challenges for missing data, as it measures a student's growth in relation to other students. As shown in Figure 2, the systematic exclusion or omission of students due to incomplete score histories could impact growth scores for all students, including those with a fully complete student record.

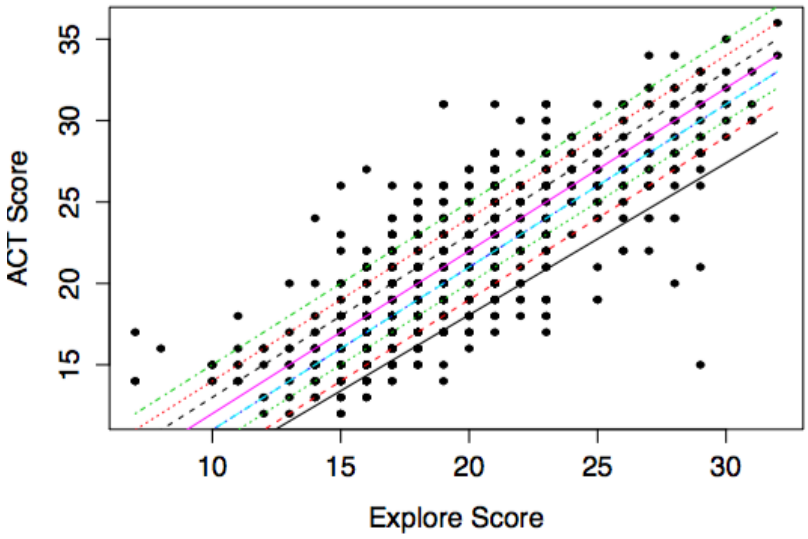
Figure 2. Changes in SGP due to Systematic Exclusion of Students



In this demonstration, excluding students due to missing data (or any other reason) could shift SGP values for the remaining students despite the same academic performance. Although this review explores missing data in the context of an SGP model, many concepts are applicable to the broader category of VAMs.

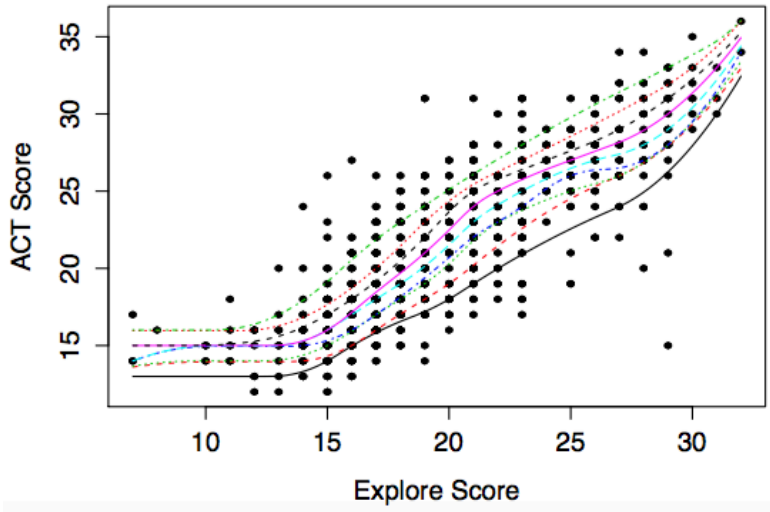
The SGP model implements quantile regression techniques to model the complex relationship between historical and future achievement trajectories. Quantile regression is similar to ordinary least squares regression, but instead of fitting the conditional mean of current scores on prior scores it fits conditional quantiles of current scores on prior scores (Koenker, 2005). Figure 3 displays deciles of ACT performance conditioned on Explore performance.

Figure 3. Quantile Regression Lines by Decile



SGP models build upon this framework by estimating quantile regression equations for the 1st through 99th percentiles. Rather than implementing a linear model, student growth percentiles are computed by fitting basis-spline, or B-spline, regression curves since educational data can be nonlinear. As demonstrated in Figure 4, nonlinearities are usually more pronounced in the low and high ends of the distribution. B-splines model the tails of the distribution more precisely.

Figure 4. Penalized Spline Model



This process is accomplished using Betebenner's "studentGrowthPercentile" function in the R programming environment (D. V. Betebenner, Adam; Domingue, Ben; Shang,Yi 2014). From this data, a matrix of scale scores and corresponding quantiles can be created for each percentile band. One strength of the SGP model is the percentile metric is easily interpreted compared to outcome measures of other VAMs. Student growth percentiles are defined as:

$$SGP = Pr(\text{Current achievement} \mid \text{Prior Achievement}) * 100 \quad (1)$$

A student's growth percentile is determined by identifying the quantile with the value closest to the student's observed score. Using the previous example, Table 1 displays a subset of possible ACT growth percentiles conditioned on Explore scores. As some assessments like ACT have discrete scale score ranges, typically the highest percentile value for each observed score is used to record a student's SGP if the same ACT score falls under multiple percentile bands.

Table 1. SGP Matrix

Explore Score	ACT score by Growth Percentile										
	P1	P10	P20	P30	P40	P50	P60	P70	P80	P90	P99
10	12	13	14	14	14	14	15	15	15	16	20
11	12	13	14	14	14	14	15	15	15	16	20
12	12	13	14	14	14	14	15	15	16	16	20
13	12	13	14	14	15	15	15	16	16	17	21
…	…	…	…	…	…	…	…	…	…	…	…
29	15	26	28	28	29	29	30	31	32	33	34
30	11	28	29	29	30	30	31	32	33	34	35
31	5	30	30	31	32	32	33	34	34	35	36
32	1	31	33	33	34	34	35	36	36	36	36

For demonstration, percentile values are displayed for deciles, though in practice additional percentile values are calculated. In this example, a student with an Explore

score of 30 and an ACT score of 32 would fall under the 70th percentile band in the matrix above. A student with an Explore score of 32 and an ACT score of 34 would receive an SGP of 50 as this is the highest percentile band for a 34 ACT score. As much of the controversy surrounding growth models lies in model inferences, SGPs offer community members, policy makers, parents, teachers, and administrators a familiar metric to base inferences. Though the SGP model was designed to provide a descriptive measure of student progress relative to their peers, these measures facilitate inferential decisions in practice, including educator evaluations.

Teacher growth scores are most commonly defined as the median growth percentile among his or her students. Some school systems define teacher growth scores as the mean SGP for his/her students, although this method is criticized because the difference between percentiles may not translate to equal growth among equally spaced percentile values. Theoretically, growth between the 50th and 55th percentile bands may be greater than growth between the 90th and 95th percentile bands. This nuance is lost when averaging SGPs.

Growth Inferences

Much of the controversy surrounding value-added modeling focuses on the inferences each type of model can support. Accurate growth interpretation is crucial, as ambiguities within growth model terminology often cloud the inferences derived from the statistical output. A fundamental challenge for VAM creators is balancing scientifically rigorous procedures (technical complexity) with easily interpretable results (transparency). This idea extends to each component of VAMs, including growth inference and missing data methodology. Complicated models may produce more

accurate results, but they will have limited utility if they are not easily communicated to educators and the general public.

Sometimes model outcomes are used to infer causality, implying teacher effects are not just attributable to a teacher but also caused by a teacher. There is substantial debate about whether or not these claims are supported by the design of various VAMs (D. W. Betebenner, 2009). The classic framework for causal inference typically includes random assignment; however value-added methodology is sometimes regarded as "an attempt to capture the virtues of a randomized experiment when one has not been conducted (Chudowsky, Koenig, & Braun, 2010)." Many school systems take this concept one step further to conceptualize projected scores as a student's performance under a typical learning environment and actual scores as the effect of their current learning environment. In this framework, a student serves as his or her own control – either intentionally or unintentionally implying causality.

In 2014 the American Statistical Association recommended increased discussion of VAM assumptions and limitations before interpreting outcome measures, specifically cautioning most VAMs quantify correlation and not causation (American Statistical Association, 2014). Further, they emphasized model limitations "are particularly relevant if VAMs are used for high-stakes purposes." The focus of this study is not whether VAMs support causal inferences but rather the effect of missing data on VAM inferences. However, sensitivity to missing data may be a consideration when discussing causality in the broader context of growth modeling.

Rubin suggests, "causal inference can be thought of as a missing data problem, with at least half of the potential outcomes missing" (D. Rubin, Stuart, & Zanutto, 2004).

Since we cannot observe the counterfactual (e.g. what would have occurred if a student participated in a different classroom or intervention), VAMs that attempt to estimate an unobserved, alternative outcome can be conceptualized as missing data models. Missing data within VAMs add an additional layer of complexity. As we continue to research the reliability and validity of VAMs, missing data methodology must be explored.

Missing Data in Growth Models

Students may have incomplete data for a variety of reasons including absenteeism, student information systems errors, inconsistent test administration, alternative testing tracks, medical emergencies, and exclusion of English-Language Learners (ELL) or Individualized Education Program (IEP) groups to name a few. For models that explicitly state how missing scores are accounted for, typically a minimum number of prior year scores are necessary to generate a predicted criterion score that is later evaluated to determine value-added growth. Some do not differentiate missing predictors (historical scores), even though not all past scores contribute equally to a student's predicted future performance. When predicting future math performance, a prior year math score is likely to carry much more predictive information than a reading score or a math score from earlier years. As a result, the pattern of missingness should inform the choice of missing data methodology.

In practice, several statewide growth models discard incomplete or partially complete student records when modeling student achievement. Records are excluded from data processing due to mismatches, out of range values, and problems with student records. Merging multiple sources of data across districts and statewide systems makes it difficult to preserve intact student records. In collaboration with American Institutes for

Research, New York state flagged missing prior year test scores and documented a greater effect for missing observations than indicators of economic disadvantage (American Institutes for Research, 2015).

A number of states including Pennsylvania and Ohio use the SASÒ Education Value-Added Assessment System (EVAAS) model to measure growth. SAS advertises a key feature that sets EVAAS apart is its ability to "[accommodate missing data] without introducing major biases by either eliminating the data for students with missing scores or by using overly simplistic imputation procedures" (SAS Insititute Inc., 2015). The SAS model, like other models, is criticized for lack of external review (Amrein-Beardsley, 2008). To date, the only study of missing data methodologies used by SAS models was conducted by its developer, potentially biasing findings (Amrein-Beardsley, 2008; S. P. Wright, 2004).

Though discussion of missing data in VAMs is limited, it is even rarer in the context of SGPs. Missing data is not formally addressed in SGP technical manuals. To date there are no routinely implemented missing data methodologies in use within SGP models, providing an opportunity for further methodological work. Though some school systems using the SGP model outline safeguards to account for missing student data in their evaluation system, most do so at the teacher level instead of the model level (Diaz-Bilello & Briggs, 2014). This ensures teachers do not receive an evaluation score based on too few observations. The danger is the SGP model is a normative growth measure, so non-random missing observations may bias the overall model used to provide a reference for each student's growth.

An analysis of student data in 2012 revealed significant differences in median growth percentiles of students eligible for free or reduced lunch and their ineligible academic peers of similar prior performance (Colorado Department of Education, 2013). These findings were replicated in Missouri when researchers used an SGP approach to model student growth (Ehlert, Koedel, Parsons, & Podgursky, 2012). Because there is variation among subgroups, if missing student observations do not adequately represent the population of students, the resulting model estimates could be biased. Assessment completion rates fluctuate by district, impacting the representativeness of the aggregate data (Brundin, 2014). These reasons and others warrant further investigation of SGP properties when modeling incomplete student observations.

Overview of Missing Data

Missing data is a frequent problem for most researchers. In theory, the best way to mitigate the consequences of missing data may be to prospectively design a study that minimizes the likelihood of incomplete observations. In practice, often the data collection process is a balance of cost, control, and feasibility that results in an imperfect final product with missing observations. Large and small-scale research projects alike are susceptible to missing data due to attrition, participant error, data collection glitches, and data entry problems. Longitudinal data utilized in student growth models is especially vulnerable to missing observations as the reasons above are compounded over multiple years, in addition to mobility in and out of the district. As there are likely unobserved covariates in every student achievement data set (e.g. student motivation), missing data methodology is relevant to all educational researchers (D. Rubin et al., 2004).

Concern about missing data is warranted given how prevalent this issue tends to be. In a review of missing data in VAMs, McCaffery found large school districts were missing at least one score from between 42 – 80% of students (D. F. McCaffrey & J. Lockwood, 2011). The distribution of missing student scores was inconsistent across teachers. On average, 37% of teacher rosters contain fully complete student records but this varies from 0 to 100% in every grade. Additionally, missing data occurred in non-random patterns that are especially relevant when selecting a missing data methodology.

Rather than discarding incomplete student records from analysis and potentially introducing bias into the sample, we wish to salvage as much data as possible to avoid loss of statistical power. Moreover, excluding students with missing data from analysis creates an issue estimating standard errors. The formula for standard error is dependent upon sample size, so reducing the sample size by even a few students adds instability to growth estimates that are then aggregated to the teacher level. This problem is particularly relevant to elementary teachers as they typically teach one class, whereas middle and high school teachers may teach several classes (P. S. Wright, 2010). Many models specify a minimum number of students that must be rostered to a teacher before an aggregated growth estimate can be calculated to avoid dramatic consequences of a reduced standard error. Still, models perform better with more students.

Fortunately, statistical packages make many missing data handling techniques readily available to researchers. Unfortunately, the most common default procedure, listwise deletion (or complete case analysis), is only appropriate for specific situations which are unverifiable and will be discussed further in subsequent sections (Peugh & Enders, 2004; Roth, 1994). This can be troubling as some researchers may not be aware

of the bias they introduce by accepting default settings. Either explicitly or implicitly, all researchers account for missing data and should be aware of the consequences of their chosen method.

Missing data methodology is a highly developed field, with seminal works produced by Rubin in the 1970s. Before that time, researchers implemented several ad hoc methods to account for missing observations. Mean imputation, regression imputation, and other single imputation procedures are still in use today, as they are easy to understand and implement despite their well-documented shortcomings. As computing technologies expanded, advanced procedures such as multiple imputation and maximum likelihood estimation came to be the preferred methods of missing data handling for most situations.

Despite these advancements, a gap remains between best practices and common practices, as ad hoc methods are still the most widely implemented procedures in educational research. Gaining attention in 1999, the American Psychological Association Task Force on Statistical Inference discouraged use of ad hoc methods, specifically referencing listwise and pairwise deletion as "among the worst methods available for practical application" (Wilkinson, 1999). Reviewing popular education and psychology journals, in 2004 Peugh and Enders found that most authors do not explicitly state how missing data was handled (Peugh & Enders, 2004). Of the studies where the missing data handling could be identified, 96% of articles employed a deletion method to account for missing observations. The remaining studies implemented either mean or regression imputation, and none used multiple imputation or maximum likelihood estimation. In 2006, Peng et. al conducted a similar review of 11 education journals and found that 97%

of identifiable data-handling techniques were deletion methods (Peng, Harwell, Liou, & Ehman, 2007). This is troubling because methodological issues from ad hoc missing data handling have been documented well before these studies were carried out.

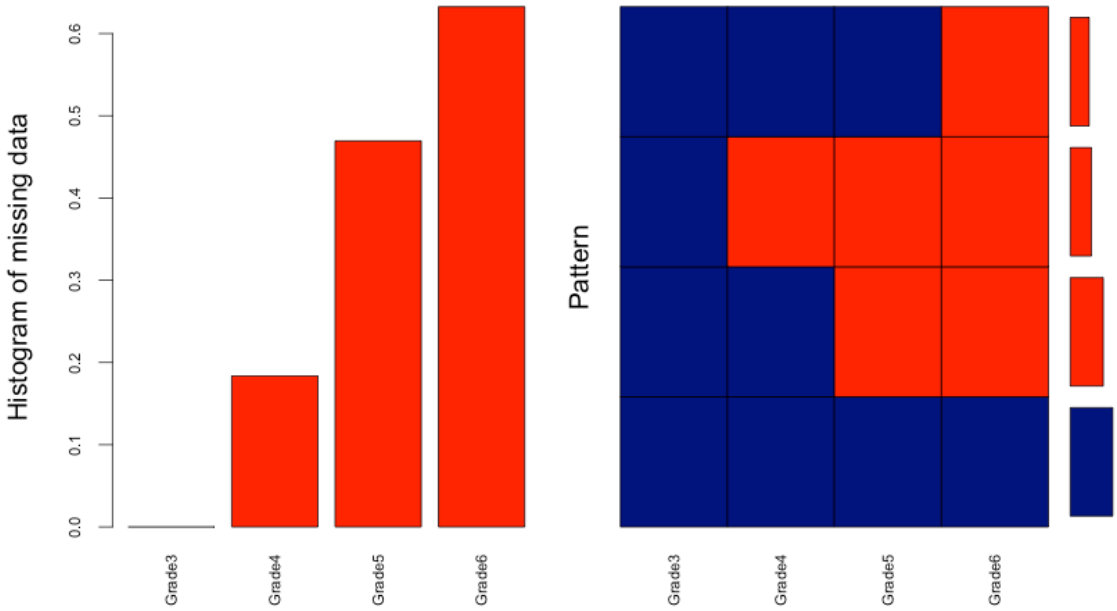
One barrier preventing applied researchers from adopting modern missing data techniques is the somewhat complicated, highly technical language of modern missing data literature. However, given the serious bias that inappropriate methods may introduce, researchers have an obligation to account for missing observations as accurately as possible. This study aims to demonstrate differences in various approaches and tie those differences to practice decisions to emphasize the impact of missing data methodologies.

Just as statisticians examine descriptive statistics of the sample before moving to analysis, it is necessary to have a sense of the amount of missing observations and patterns of missingness present in the data before deciding how to account for missing data. Patterns may shed light on the missing data mechanism or highlight errors in data collection that can be corrected. Exploratory analyses of missing observations usually include the percent and frequency of missing observations, and whether or not missing values are clustered among variables. The more that is known about missing values, the more confident the researcher can be in the choice of missing data method (Honaker, King, & Blackwell).

Some missing data methods perform better when monotone missingness patterns are observed, particularly those that model the missingness in conjunction with the outcome measure (Carpenter & Kenward, 2012). A monotone pattern exists when missing observations for a particular variable are always missing in subsequent

observations. Figure 5 illustrates the frequency (left panel) and pattern (right panel) of missing observations.

Figure 5. Missing Data Patterns



The histogram shows no data are missing in Grade 3, though the percent of missing data increases in grades 4 through 6. The pattern of missing observations displayed in the right panel of Figure 3 show these values are missing in a monotone fashion. In the pattern plot (right panel), blue represents complete data and red represent missing or incomplete data. In this demonstration, all 3rd grade scores are complete, indicated by all blue squares in the bottom row of the pattern plot. The next row shows complete scores for 3rd and 4th grade, but missing scores for 5th and 6th grade. Remaining rows show complete data for third grade only, and then complete data for grades 3-5 but missing in grade 6. This scenario qualifies as a monotone pattern because cases with missing values at a given point in time are also missing values for subsequent observations. Monotone

patterns common for longitudinal studies (due to attrition) and in survey research (when participants decide to stop).

Enders defines a general missing data pattern as "missing values dispersed throughout the data matrix in a haphazard fashion" (Enders, 2012). However, he cautions that the patterns of missingness should not signal causality, in that the reasons for missingness may not be random even if the pattern appears so.

Missing Data Mechanisms

Rubin's taxonomy of missing data mechanisms has become the standard classification scheme cited in most research (Donald B. Rubin & Wiley, 1987). He specified three mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). To demonstrate each condition, Table 2 presents all three missing data mechanisms imposed on fictitious student test data. Free/reduced lunch participation is also included, as it commonly serves as a proxy for socioeconomic status in educational research.

Table 2. Student Test Scores with MCAR, MAR, and MNAR Mechanisms

Free/Reduced Lunch Participation	Complete	MCAR	MAR	MNAR
No	99	99	99	99
No	95	--	95	95
No	95	95	95	95
No	92	--	92	92
Yes	90	90	--	90
No	89	89	89	89
No	85	85	85	85
Yes	76	76	--	--
No	67	67	67	--
Yes	55	--	--	--

Data are MCAR when the probability of missing observations is independent of any other variable (latent or observed). Essentially, data are arbitrarily missing and thus the observed data can be considered a random sample of the complete dataset. In the example in Table 1, missing test scores in the MCAR condition are scattered randomly and are unrelated to free/reduced lunch participation or the test score itself. This situation is ideal as it lends itself to the most methods to account for missing data. Though estimates derived from a MCAR dataset will largely be unbiased if missing data are omitted from analysis, the main drawback of a MCAR mechanism is loss of statistical power.

Data are MAR when the probability of missing observations is independent of the missing variable itself, but related to another variable. In the example in Table 2, missing observations in the MAR conditions are not a function of test scores, but are related to free/reduced lunch participation. As the missingness is conditional on another variable in the dataset, there are a variety of methods available to restore attributes of the would-be complete dataset using information from other non-missing variables. More relaxed than the MCAR condition, most missing data procedures require data to be MAR. There are no formal diagnostic tests to detect a MAR mechanism.

Data are MNAR when the probability of missing observations is a function of the missing variable itself. In the example in Table 2, all test scores below 85 are missing. Missing data are said to be "non-ignorable" or "inaccessible" if the missing data mechanism is MNAR. This mechanism is most problematic for researchers, and requires specific analysis techniques (e.g. selection models, pattern mixture models) that are beyond the scope of this study.

Missing data mechanisms apply both to individual data points as well as the analysis. For example, outcome variable Y is predicted by student test scores (X1) conditioned on free and reduced lunch participation (X2). Assuming missing test score data is attributable to free and reduced lunch participation, this analysis would fit a MAR mechanism as long as both X1 and X2 are included in the model. However, a model where X1 is the only predictor of Y (not conditioned on X2, the cause of the missingness) may be defined as MNAR (Graham, 2009). As all three mechanisms can exist concurrently within the same dataset, at best we can "make plausible guesses about [their] relative contributions and examine the probable effect of inaccessible missingness given a range of plausible assumptions (Graham, Taylor, & Cumsille, 2001)." Because missingness mechanisms cannot be verified, statisticians can conduct sensitivity analyses assuming different mechanisms to determine how robust their findings are.

Missing Data Methodologies

After considering the missing data pattern and mechanism, there are countless methods available to analyze data sets with incomplete observations. This review focuses on common techniques applicable to VAMs. The goal of implementing any missing data handling technique should be to produce unbiased parameter estimates with accurate variability (e.g. standard error) while retaining as much statistical power as possible. Most fall under deletion methods, imputation methods, or likelihood-based methods and have tradeoffs in terms of assumptions and efficiency.

Methods specific to MNAR contexts are omitted from this discussion for several reasons. Situating missing data methods within a framework of missingness mechanisms is an important thought exercise, but the practical implementation of these procedures is

less straightforward given mechanisms are unverifiable and can occur simultaneously. MNAR is the most extreme of the 3 mechanisms, potentially limiting its application. Most VAMs assume MAR, and teacher effects assuming MAR and MNAR have been shown to be similar (D. F. McCaffrey & J. R. Lockwood, 2011). Last, some methods presented below do not require the researcher to make any assumptions about missing values.

Deletion Methods

Listwise deletion, or complete case analysis, is the default missing data handling approach for most statistical packages and is the most popular method cited in educational research (Peugh & Enders, 2004). Only cases with full information are included in the analysis. In the context of student growth models, a listwise deletion approach would eliminate any student without a complete set of historical achievement scores from the model. Though this approach is attractive because it requires no additional computations to account for missing observations, its major drawback is that it can result in biased parameter estimates unless data are MCAR. Allison warns if the data vary across subgroups, "any nonrandom restriction of the sample (e.g. through listwise deletion) may weight the regression coefficients toward one subset or another (Allison, 2002)."

Even assuming MCAR, this approach is not preferable as it results in a smaller effective sample with reduced statistical power. At worst, using a data set of 1,000 observations across 5 variables with 5% missing values for each variable, the effective sample size is reduced to 774, as .955 x 1,000 = 774. These calculations assume unique cases are only missing one variable. Similarly, using a data set of 1,000 observations

across 5 variables with 20% missing values for each variable, the effective sample size is reduced to 328. If cases have multiple variables with missing values, the effective sample sizes will be larger as fewer cases are omitted per missing value.

Pairwise deletion, or available case analysis, salvages more data than listwise deletion by estimating correlations between variables using as many observed cases as possible. This allows a correlation matrix to be computed with different sample sizes for different combinations of variables. The SPSS statistical package is "by far the most dominant package" cited in journal articles today (Muenchen, 2015). SPSS defaults to pairwise deletion when producing correlations and allows pairwise deletion as an option for other analyses as well. Pairwise deletion is detectible when published studies produce different sample sizes for different procedures conducted on the same data set (Enders, 2012). A key problem with this approach is that a standard error cannot be accurately estimated as sample size is part of the equation, and sample size varies depending on the parameter estimated. As with listwise deletion, this method is only appropriate for MCAR conditions because if observed data differ systematically from missing data, estimates derived from available case analysis may be biased (Gelman & Hill, 2006).

Imputation Methods

Missing data approaches that estimate either individual missing values or distributions of plausible missing values all fall under the category of imputation methods. Unlike deletion methods, these techniques retain all of the data collected. The imputation process prepares data for the substantive primary analysis, giving the researcher flexibility to conduct a broad range of post-imputation analyses since missing

data are accounted for in the imputation phase. However, substantive analyses are contingent upon proper specification of the imputation model.

Among the simplest imputation models widely used in practice is mean imputation. Mean imputation, or mean substitution, replaces missing values with the variable mean, and has been a popular method for estimating missing observations because it is conceptually simple and easy to implement. Imputing the mean salvages incomplete data and will not distort mean estimates. However these benefits are offset by its deficiencies. Imputing missing values with the average value for that variable will constrict its variance as well as its covariance with other variables. This method is not suited for any missing data mechanism. Often dummy variable adjustment is used in conjunction with mean imputation. After dummy coding imputed cases (e.g. complete cases=0 and imputed cases=1) and regressing a dependent variable on a set of independent variables, in theory, variation attributable to missing observations should be accounted for, however this method still results in variance underestimation (Cohen, Cohen, West, & Aiken, 2013). However, this method generally results in biased parameter estimates even when the data are MCAR. Under no circumstances (MCAR, MAR, or MNAR) is this method appropriate (Allison, 2002).

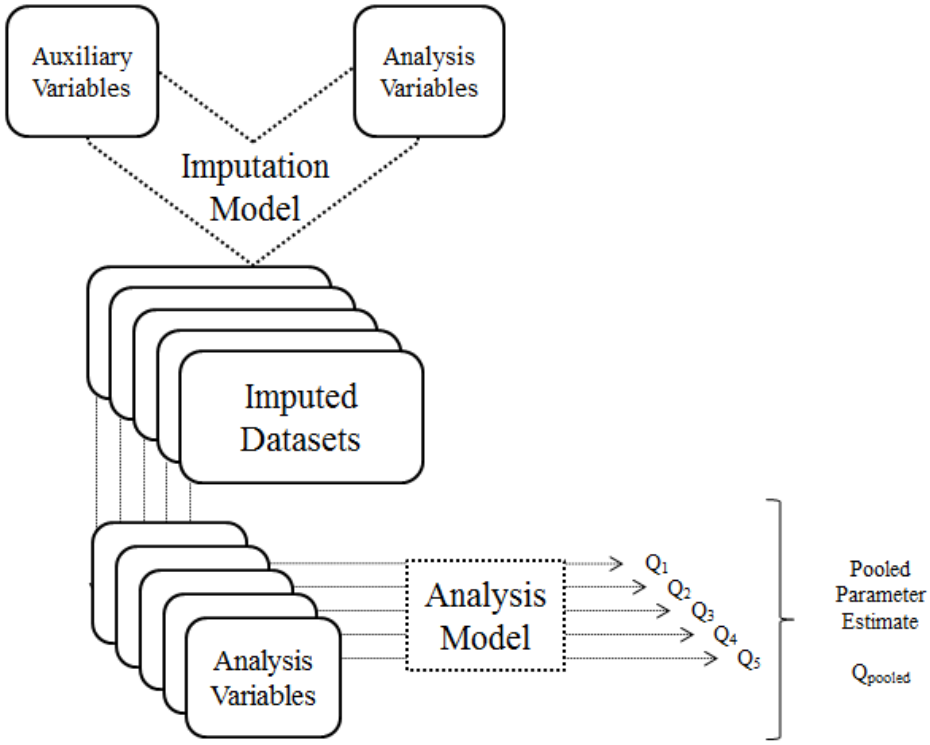
Conditional mean imputation, or regression imputation, replaces a missing value with the average value conditioned on other observed variables in the form of a regression equation. Because all missing observations will be imputed with values on the regression line, imputed cases have no residual variance. This process shrinks the standard error and overestimates the association between the variable with missing observations and other variables in the model. If the researcher has information about

what data are missing, weighted least squares regression (weighted on population distributions to correct for a disproportionate observed sample) may improve parameter estimates. Though weighting can compensate for a systematic exclusion of subpopulations, it requires the researcher to specify the population distribution and assumes subpopulations responses are representative (e.g. no response bias causing missing observations).

Another improvement on this process is stochastic regression, which adds randomly distributed residual error to each imputed case. Stochastic regression is the only single imputation technique found to produce unbiased parameter estimates under MAR conditions (Enders, 2012). Adding error may seem counterintuitive at first, but this approach works because the focus is not accurate predicted values for individual cases. Instead, imputing data points with error preserves the variability of the would-be complete dataset resulting in accurate parameter estimates.

Whereas single imputation methods estimate individual values for missing observations that are then treated as observed values in the analysis, multiple imputation (MI) techniques estimate a distribution of plausible values. The purpose of MI is not to estimate individual data points, but instead to preserve properties of the would-be complete dataset had there been no missing observations (Enders, 2012). Building on the underlying principles of stochastic regression and Bayesian estimation, MI generates multiple plausible values for each missing observation and then pools estimates derived from the primary analysis of interest (in this study, growth estimates), as demonstrated in Figure 6. MI is performed iteratively until the specified number of imputations is met or convergence is reached.

Figure 6. Multiple Imputation Process



There are several imputation algorithms available, but the most widely used method in many statistical packages is described here. First, stochastic regression is used to generate regression equations for imputation. The Markov Chain Monte Carlo (MCMC) technique first simulates a random independent draw from the conditional distribution of missing values given the observed. Next, Bayesian methodology is used to estimate parameter values of the posterior distribution. These values are then used to inform repeated-imputation inference (Schafer, 1999). Each iteration refines the estimation as the mean and covariance vectors from the previous imputation are used to construct new regression equations for the next round of imputation. Random draws of residual error are added to each element. This process is an MCMC procedure as the

initial estimate is a random imputation, thus a Monte Carlo technique, and each subsequent step is dependent only on the previous step, thus a Markov Chain, defined by the following:

$$(Y_{mis}^{(1)}, \theta^{(1)}),\ (Y_{mis}^{(2)}, \theta^{(2)}), \ldots (Y_{mis}^{(k)}, \theta^{(k)}), \quad (2)$$

where Y_{mis} represents missing observations and θ represents the current parameter estimate of interest (Liang, Liu, & Carroll, 2011). Several datasets are generated with different parameter estimates. Finally, the results are pooled into a single set of parameter estimates that reflect our uncertainty about missing observations and ordinary variability among samples.

An alternative method for multiple imputation uses a predictive mean matching approach to match observed and missing cases before imputing missing data. Predictive mean matching (PMM) is an alternative, semi-parametric imputation approach. Missing observations are replaced with "donor" values in the form of the closest observed value to the regression-predicted score (Landerman, Land, & Pieper, 1997). To impute m number of imputations, the researcher imputes *m*=*n* random draws from the k donor values closest to the predicted value. There are several methods for specifying the donor pool, including 1) a fixed approach where k possible donor values are specified (e.g. *k*=5); 2) a caliper matching approach defined by a specified caliper width; and 3) an approach that sets the number of donor values to the number of observed values, but assigns closer donor values a higher probability of selection.

Unlike other imputation algorithms, all imputed values are within the observed score range since they must come from other cases in the data set. As the number of

missing observations increases (and subsequently the pool of donor values increases), the variability of point estimates increases (Morris, White, & Royston, 2014). A potential pitfall for PMM is "donor sparseness," when few donors are similar to predicted values. However, the main advantage of PMM is this approach is robust to violations of joint-normality assumptions required by other parametric imputation procedures.

While MI applications are readily accessible to researchers through popular statistical packages, this process cannot be automated to the extent that the researcher need not make decisions for each specific application. Common misspecifications include omitting the outcome variable from the imputation model, and improperly modeling non-normal distributions (Sterne et al., 2009). With every imputation strategy, the imputation model must be compatible with the analysis model. Complicated analyses require equally complex imputation models to preserve attributes of the would-be complete distribution. If the analysis model includes squared terms, interactions, or other transformations, the imputation model must include the same terms as not to impute bias into the model (Carpenter & Kenward, 2012). Though this sounds intuitive, the analysis model may not be clearly defined before the imputation model is developed.

Specifying an imputation model is as much art as it is science. Often imputation models are constructed with a large number of predictor variables to utilize as much information as possible. Overly simple imputation models may downwardly bias correlation estimates in the analysis model if they do not adequately capture dependencies between variables in the imputation model (Sterne et al., 2009). Other situations lend themselves to more parsimonious, slimmer models. Since MAR is an assumption rather than an attribute of the data, the researcher must determine whether or

not this assumption is satisfied and if MI is appropriate. No imputation approach is appropriate for every context, requiring researchers to carefully consider each phase of model specification.

Some researchers face resistance when implementing imputation methods, fighting the perception they "make up" data points to benefit their hypothesis (Wayman, 2003). This perspective overlooks the main goal of imputation: to recover attributes of the complete sample. All data is measured with error and models routinely make predictions that could be conceived as "made up" estimates of future performance. Even complete test score data carries error when students repeat an exam and receive different scores. Similar to test-retest reliability, statisticians largely regard imputation procedures as a routine element of statistical practice (Fichman & Cummings, 2003). Still, communicating model results to a wider audience may be challenging if data are imputed, evident in communication from the Pennsylvania Department of Education (PDOE) highlighting statistical properties of the Pennsylvania Value-Added Assessment System (PVAAS). In a document debunking common misconceptions about PVAAS, the PDOE explains their model accounts for missing data without imputation techniques so that "…no values are explicitly imputed (statistically "made up") for the missing scores! (PDE, 2015)"

Likelihood-based Methods

Maximum Likelihood (ML) is a procedure to estimate parameters by finding the most probable values for missing observations given observed distributions. By retaining information about other variables, ML techniques improve model accuracy by "borrowing" information from observed attributes to estimate missing attributes (Enders,

2012). As each variable relates to other variables in the model, improving the predictive ability of one variable can affect parameter estimates of other variables. Assuming reading, science, and math scores have a joint normal distribution and are all correlated, missing reading test scores can be estimated by borrowing information from math and science scores. Different "auditions" are compared by their log-likelihood values, with the smallest value indicating the highest likelihood (Enders, 2012). Log-likelihood computations use only complete data, creating a different formula for each missing data pattern. Mixed models employ a restricted maximum likelihood technique that accounts for missing observations.

As computing technology advanced, MI and ML grew in popularity simultaneously. However, they should not be framed as competing approaches; each produce consistent estimates when implemented appropriately. Both procedures require multivariate normal data and MAR assumption, and produce asymptotically equivalent results as sample sizes increase (Enders, 2012). Though they share attractive properties, MI and ML methodologies are fundamentally different. The MI process relies on posterior probability. Posterior probability is the opposite of maximum likelihood in that it represents the probability of parameter estimates occurring given evidence from the data, whereas ML techniques represent the probability of observed outcomes occurring given parameter estimates. Unlike MI, traditional likelihood-based models account for missing data and perform the analysis of interest simultaneously, meaning ML methods only use information from variables included in the analysis (no auxiliary variables). Another drawback of maximum likelihood based methods is they cannot accommodate outcome-dependent missingness. In contrast, multiple imputation procedures can impute

missing data among the outcome variable as well as use outcome variable information as a predictor when imputing missing data in other variables.

Inverse Probability Weighting

Again, the more information the researcher has about missing data, the easier it is to account for missing observations. In situations where the researcher has prior information about a population distribution, inverse probability weighting (IPW) can be useful to upweight subgroups that are underrepresented due to missing data (Seaman & White, 2013). Similar to selection sampling techniques, IPW re-weights data to create a desired pseudo-population that mirrors known population characteristics. IPW is a complete case analysis that weights cases by the inverse of their probability of being complete. For example, if school district enrollment documents a known percentage of low-income students, but student achievement data is only available for a subset of this group, IPW can account for the discrepancy between observed and missing student achievement scores by assigning a greater weight for low-income students. Unlike MI and ML, this approach requires the researcher to specify a model for the probability of missingness but makes no assumptions about the analysis model. Therefore it is not limited to a joint normal distribution. Similar to MI, analyses incorporating IPW occur in two phases. This allows the researcher to take advantage of information provided by auxiliary variables when specifying the missingness model. The inverse of the predicted probabilities for a complete record then provide analysis weights.

If the probabilities of missingness are accurately accounted for, IPW estimators are consistent regardless of the mechanism (MAR, and MNAR). Weights are generally more precise in larger samples, though corrections can be used for small sample

estimation (e.g. SAS PROC G). There are also methods to stabilize weights (Carpenter & Kenward, 2012). While previous literature documents the inefficiencies of IPW compared to MI and ML, it trades efficiency for robustness. IPW requires fewer assumptions and can be applied to a variety of circumstances.

CHAPTER THREE

METHODS

As described in detail in chapter 2, VAMs present methodological challenges in regards to missing data. Therefore missing data methodologies are explored in the context of the Student Growth Percentiles (SGP) model. It is not a goal of this analysis to document the validity of the SGP model for measuring student achievement growth or educator effectiveness. Instead, the primary goal of this analysis is to quantify the variability in educator evaluations due to missing student data. The following research questions are addressed:

1. How sensitive are growth estimates to missing data?
2. Does the choice of missing data methodology result in different growth inferences when used in an educator evaluation framework?

To assess the adequacy of each methodology in recovering Student Growth Percentiles (SGPs), Median Growth Percentiles (MGPs), and teacher proficiency ratings, estimates were compared using the following: listwise deletion, likelihood-based imputation using an Expectation-Maximization (EM) algorithm, multiple imputation using a Markov Chain Monte Carlo (MCMC) method, multiple imputation using a Predictive Mean Matching (PMM) method, and Inverse Probability Weighting (IPW). Additionally, this study explores how results of each approach translate to evaluation inferences.

SGP Model Specification

Before artificially censoring observations, SGP estimates were calculated and aggregated to the teacher level to serve as a benchmark for comparing missing data methods. Five copies of the censored dataset were used to impose each of the 5 missing data methods in this study. After preprocessing data using each missing data method, growth was calculated using a student growth percentiles model. No demographic characteristics are used in the growth analysis; 3rd and 4th grade mathematics scores are the only variables used in the SGP model. To generate student growth estimates, the τ^{th} quantile of 4th grade mathematics achievement, represented as $Q(\tau|\, X = x) = x'_i\beta(\tau)$ is solved by the following:

$$\hat{\beta}(\tau) = arg\min_{\beta \in R^P} \sum_{i=1}^{n} P_\tau(y_i - x'_i\beta) \qquad (3)$$

where $0 < \tau < 1$ (Chen, 2005). This means $\tau = .25$ represents the 25th percentile, $\tau = .5$ represents the median or 50th percentile, and $\tau = .75$ represents the 75th percentile. The SGP model estimates quantiles 1 through 99 so comparisons to fitted values can be made. A student's SGP was determined by the closest quantile curve to their actual score given their prior test history.

The SGP model implemented in this study sets four interior knots and two boundaries. Regression quantiles are estimated for quantiles 1 to 99 given current and prior achievement, however not all quantiles may be observed when calculating SGP scores. For example, if only two students received a 3rd grade baseline score of 180, a maximum of two distinct SGPs would be observed for this part of the conditional distribution though quantiles 1 through 99 are estimated. SGPs are generated comparing predicted values based on a prior year score to expected values to find the closest

percentile value. As shown in Table 3, a student that had a 3rd Grade Score of 240 and a 4th Grade Score of 255 would receive an SGP of 75 as this percentile is the closest predicted value.

Table 3. Example SGP Matrix

3rd Grade Score	4th Grade Mathematics Percentile				
	P5	P25	P50	P75	P95
160	160	160	160	170	185
171	160	165	166	175	225
…	…	…	…	…	…
239	240	242	245	250	250
240	240	245	250	255	260

A student that received both a 160 for both 3rd and 4th grade would receive an SGP of 50, as this is the highest of several quantiles that produce the same predicted score value. Estimated quantiles that are close in their predicted value for 4th grade scores are noticeable in Figure 9 (below) in areas of the distribution where quantile regression curves are either close in proximity or overlapping.

Sample and Data

To illustrate the consequences of different missing data handling techniques for student growth data, this study analyzes Measures of Academic Progress (MAP) mathematics achievement scores. Actual test scores were chosen over simulated data to ensure the relationship between past and future performance accurately reflects what exists in practice. Though simulating scores could provide additional statistical control over missing data patterns/mechanisms, teacher effect sizes, and other variables, these controls may not translate to practice settings where data is often more complicated.

In general, more test score data usually translates to greater predictive information. Because the SGP model can accommodate several prior years of data and

not all missing observations contribute equally to a student's predicted future performance, arguably teachers of later grades may be less affected by missing observations than their counterparts who teach elementary grades with less historical data. To isolate the effect of missing data from confounding variables (e.g. different number of predictors, different measurement error across assessments, etc.), this analysis concentrates on evaluations across a single subject and grade. For an evaluation scenario with one prior year of data, 4th grade mathematics growth was evaluated using 3rd grade math achievement as the single predictor. Since this dataset contains missing data like all large educational datasets, missing observations were removed to arrive at a pseudo-population of complete scores for analysis. Analysis of the full set of 415 students with all scores in tact serves as the basis for comparison for missing data methods and is hereafter referred to as the benchmark analysis.

Mechanism to Assign Censoring

Since missingness mechanisms cannot be verified to inform our choice of missing data methodology, a reference population was used to simulate a pattern of incomplete records instead of specifying a missingness mechanism. Missing student data occur for a variety of reasons, perhaps because multiple missingness mechanisms are at work simultaneously. Patterns of missingness observed in from 6th grade student records outside the analytic sample were used to inform the missingness mechanism imposed on the analysis sample of 4th grade students. This reflects our uncertainty about why student records are missing yet mirrors complex patterns that occur in practice. Propensity score matching of complete and incomplete student records in the reference population was conducted using the following variables: IEP status, LEP status, gender, ethnicity, free or

reduced lunch eligibility, and a standardized prior mathematics achievement score. Equations generated were used to artificially censor 4th grade scores in the analysis sample. As students are non-randomly assigned to teachers, student rosters were not manipulated. It is important to note that student growth is the outcome of interest in the primary analysis model (the SGP model) used in this study. Growth was calculated using 4th and 3rd grade mathematics scores only; demographic variables were not included in the SGP analysis.

Benchmark Analysis and Listwise Deletion

Both the benchmark (pre-censored) analysis that serves as a comparison group for all other missing data methodologies and the listwise deletion analysis do not require any additional computations to preprocess data before conducting the SGP analysis. Since listwise deletion is itself a complete cases analysis, these two models were identical in all but their inputs. As described in subsequent sections, other missing data procedures account for missing data before implementing the same SGP analysis.

Likelihood-based Imputation using an EM Algorithm

The EM algorithm does not impute scores directly. Instead, it estimates parameters by maximizing the observed data log likelihood function iteratively, using an E-step and M-step. Using Q to denote the statistic of interest, the probability of Q given observed values Y_{obs} can be expressed using the following:

$$P(Y|Q) = P(Y_{obs}, Y_{mis}|Q) = P(Y_{obs}|Q)P(Y_{mis}|Y_{obs}, Q), \quad (4)$$

The log likelihood function of the above equation is:

$$l(Q|Y) = l(Q|Y_{obs}) + \log P(Y_{mis}|Y_{obs}, Q) \quad (5)$$

To estimate Y_{mis}, this technique first calculates model parameters (e.g. means, variances, and covariances) given the complete data. Maximum likelihood techniques produce regression equations for each variable given its relationship with observed variables. These equations are then used to produce estimates for missing observations. Using the newly complete dataset of imputed and observed scores, parameter estimates are re-calculated as more data is available. This process is repeated iteratively until convergence is reached. To reflect our uncertainty about the missing observations, normally-distributed stochastic error is introduced to parameter estimates. For application in this study, these resulting parameter estimates were used to impute individual missing values through linear regression. Maximum likelihood estimates of the mean vector and covariance matrix were obtained at the final iteration and were used to estimate single values for each missing 4th grade mathematics score. The newly "complete" data set was then used for the computation of student growth percentiles.

Multiple Imputation by Markov Chain Monte Carlo (MCMC)

This technique uses a stochastic model to produce m number of imputations with m corresponding parameter estimates, reflecting our uncertainty of the missing observations. Through an imputation step (i-step) and posterior step (p-step), MCMC iterates between likely imputation values and the resulting posterior distribution. Again, using Q to denote the statistic of interest, the probability of Q given observed values Y_{obs} is expressed using the following:

$$P(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis})P(Y_{mis}|Y_{obs})dY_{mis} \quad (6)$$

Next, the I-step is expressed as:

$$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} \middle| Y_{obs}, Q^{(t)}, U^{(t)}\right), \tag{7}$$

where *U* is the estimated variance of *Q*, and t is an indicator of time used to order each step. In this study a non-informative prior was used. Once the first set of estimates were computed, new values were drawn from the posterior distribution of the newly complete dataset (using observed and imputed values). The p-step is expressed by the following:

$$Q^{(t+1)}, U^{(t+1)} \sim P(Q, U | Y_{obs}, Y_{mis}^{(t+1)}) \tag{8}$$

The posterior distribution of *Q* is an average of repeated draws from $P(Y_{mis}|Y_{obs})$, posterior predictive distributions of missing data given observed data.

The analysis produced five imputed datasets (*m*=5) to analyze separately using the SGP model to calculate student growth percentiles. After *m*=5 imputations were calculated, SGPs were converted to normal curve equivalents (NCEs) for pooling using the following equation:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \widehat{Q}_i \tag{9}$$

This was necessary since the SGP metric is not suitable for pooling given it is not of an equal interval property. Pooled NCEs were then converted back to the SGP metric to arrive at the final estimates used for growth inferences and evaluation purposes.

Multiple Imputation by Predictive Mean Matching (PMM)

This method is similar to the linear regression-based MCMC in its methodology but generates final imputation values from donor values of similar observed scores. First, regression equations are used to estimate parameters given observed values. Parameter

estimates are used to create predicted values for both observed and missing values. A pool of potential donor values (*k*) is formed using the closest observed predicted values to each predicted missing value. In this study, *k* was set to 1 (the default value for SPSS and the mi command in Stata) meaning predicted values of missing cases are matched to the observed case with the closest predicted value. Imputations are random draws from the donor pool and are imputations are matched to missing observations via their predicted values. In this study, imputations derived from a PMM algorithm were calculated using the following:

1. Obtain $\hat{Y}_{obs}$= $\{\hat{Y}i = x_i^T\beta : i \in Observed\}$
2. Obtain $\hat{Y}_{mis} = \{\hat{Y}j = x_i^T\beta : j \in Missing, i \in Observed\}$
3. Locate $\hat{Y}_{obs}$ observations with predicted values closest to $\hat{Y}j$ for all $j \in Missing$.
4. For *m*=*n*, impute random draws from *k* observations from donors closest to the predicted values of $\hat{Y}_{mis}$.

Last, similar to other MI approaches, after analyzing imputed datasets separately to produce different sets of SGPs, SGP estimates were converted to normal curve equivalents (NCEs) and pooled. Pooled NCEs were then converted back to the SGP metric.

Inverse Probability Weighting

Unlike the other methods explored in this study, inverse probability weighting (IPW) methods re-weight data to create a pseudo-sample that mirrors known population characteristics. IPW methods do not estimate or impute missing observations, and are a complete case analysis similar to listwise deletion in that they omit any cases with incomplete data. Instead, IPW procedures upweight cases that may be underrepresented

due to missing data. Similarly, overrepresented student groups are assigned a lower weight to make the observed sample more proportionate to the population.

First, in order to derive probabilities necessary to calculate analysis weights, logistic regression was used to model complete or incomplete record status. An indicator variable, R_i, denotes a fully complete student record. Inverse probability weighting methods weight the i^{th} observation by R_i / π_{i0} (the inverse of its probability of being observed). For example, a student with complete data and $\pi_{i0} = .2$ is given the weight of five students in an attempt to make the sample more representative of the would-be complete population.

In this study, the probability of inclusion was determined by modeling missingness in a reference population of 6th grade students outside the analytic sample of 4th grade students. The following terms were used as predictors in the logistic regression model: gender, ethnicity, LEP status, IEP status, free or reduced lunch eligibility, and standardized prior year mathematics achievement score. Coefficients derived from the missingness model were then used to generate a predicted probability of inclusion for each student in the analytic sample of 4th grade students.

Next the primary analysis of interest, the SGP analysis, was carried out using the weighted dataset to produce quantile estimates of 4th grade growth from 3rd grade baseline mathematics achievement. After estimating growth quantiles using the weighted data, a student's SGP was determined by identifying the quantile with closest predicted value to the student's actual 4th grade score. The process of assigning an SGP remains the same for IPW data as it was for all other study scenarios; a matrix of SGP quantiles was used as a lookup table to identify the closest expected and actual 4th grade score

values given a student's 3rd grade achievement score.

Weights were used to create a pseudo-sample representative of the would-be population in creating the SGP matrix, however they were not used to assign a growth score. The SGP growth metric itself was not weighted in subsequent calculations, meaning each student received one growth score and no student's SGP is weighted more than any other student's. The goal of IPW was to rebalance an unrepresentative sample of students when generating quantile estimates. Thus, weights were only utilized to generate a matrix of SGP quantiles and corresponding predicted scores; weights were not used in any other calculation (e.g. aggregating growth scores).

Evaluation Classification Methods

Part of any successful data analysis is extracting meaning from a dataset. Evaluation inferences derived from VAMs are the most controversial aspect of growth modeling; therefore, documenting changes among teacher proficiency categories is essential. The frequency of misclassifications within the VDOE evaluation framework and overall magnitude of rating bias will be explored. This element grounds the methodological findings, as differential model precision may not be relevant to practice. Conversely, seemingly negligible differences in parameter estimates could translate into unacceptable inference fluctuations given the high-stakes environment of teacher evaluations.

Thresholds were needed to determine the magnitude of teacher evaluation misclassifications that are attributable to missing data, so the approach by the Virginia Department of Education (VDOE) was adopted. The extent to which these proficiency categories represent differences in educational effectiveness is beyond the scope of this

study. Rather, documenting movement among these categories is intended to demonstrate the practical consequences of missing data on teacher evaluations. VDOE uses the following framework to determine teacher evaluation scores (Jonas):

Table 4. Growth Classifications

Student Growth Categories	
Low growth	SGPs of 1 to 34
Moderate growth	SGPs of 35 to 65
High growth	SGPs of 66 to 99
Teacher Evaluation Categories	
Exemplary	More than 50 percent of students demonstrated high growth and no more than 10 percent demonstrated low growth
Proficient	At least 65 percent of students demonstrated moderate or high relative growth (the percentage of students with high growth + moderate growth > 65 percent)
Developing/ Needs Improvement	< 65 percent of students demonstrated moderate or high growth; AND < 50 percent of students demonstrated low growth.
Unacceptable	> 50 percent of students demonstrated low growth

The VDOE framework is one of many evaluation frameworks used in practice. Others use the Median Growth Percentile metric for educator evaluation purposes. To consider the impact of each missing data method on MGP estimates used for accountability purposes, the Massachusetts Department of Elementary & Secondary Education (MDOE) framework was also considered. MGPs below or equal to 35 are categorized as low, MGPs above or equal to 65 are categorized as high, and MGPs between 35 and 65 are categorized as moderate.

Criteria for Comparisons

Since missing data techniques have different goals (producing unbiased parameter estimates, estimating accurate variability, and retaining statistical power), several

methodological properties were explored. Correlations between MGP estimates for each condition are compared. Rank correlation coefficients between MGPs and complete/incomplete student observations statuses (frequency of student missingness) were calculated. Mann-Whitney tests were used to detect differences in MGP distributions derived from each missing data methodology compared to the complete distribution. To compare the magnitude of growth differences, mean absolute errors in growth percentiles were compared.

CHAPTER FOUR

RESULTS

The results of this study are organized by the two overarching research questions, first exploring the methodological impact of missing data in estimating student growth, and then considering practical implications for implementing missing data methods in practice. Chapter V discusses findings from both research questions as they relate to education policy.

Research Question 1

How sensitive are growth estimates to missing data?

Data

Students in this study consist of a single cohort of 4th grade public school students from one school (n=415) over two academic years. Growth from 3rd to 4th grade was evaluated using Student Growth Percentile (SGP) methodology. Assessment data include 3rd and 4th grade Measures of Academic Progress® (MAP) mathematics achievement scores. The sample means for 3rd and 4th grade MAP scores were 205.4 and 213.8 respectively, slightly higher than the national norms estimated by the publisher (203.1 for 3rd grade and 212.5 for 4th grade). Classroom rosters were used to link mathematics teachers to students as part of an accountability framework. In addition to student achievement scores, data include the following demographic characteristics: Free and Reduced Lunch eligibility, LEP status, IEP status, Gender, and Ethnicity.

Descriptive statistics for student demographic characteristics are provided in subsequent sections.

Assigning Artificial Missingness

A reference population of 6th grade students from the same school was used to explore missing data patterns found in practice. Propensity scores were calculated by regressing an indicator variable for a complete or incomplete student record (1=complete, 0=incomplete) on student assessment history and demographic characteristics. While propensity score methodology is commonly used to process quasi-experimental data for causal inference, its primary function in this study is to distill a multidimensional covariate profile into a single dimension. By matching students on demographic and achievement variables, propensity scores allow comparisons to be made between students of similar propensity scores while retaining information from multiple dimensions used in their calculation.

Propensity scores were calculated for 4th grade students in the study sample using the regression coefficients derived from the reference population. Students with the lowest predicted probabilities for complete data were flagged for censored status, and their 4th grade mathematics scores were removed from the analytic sample. This step was implemented to adjust for differential probabilities of observing complete data among various student groups when assigning artificial missingness in the analysis sample. Figures 7 and 8 present student characteristics and their associated predicted probabilities for complete data.

Figure 7. Propensity Scores Disaggregated by Student Ethnicity

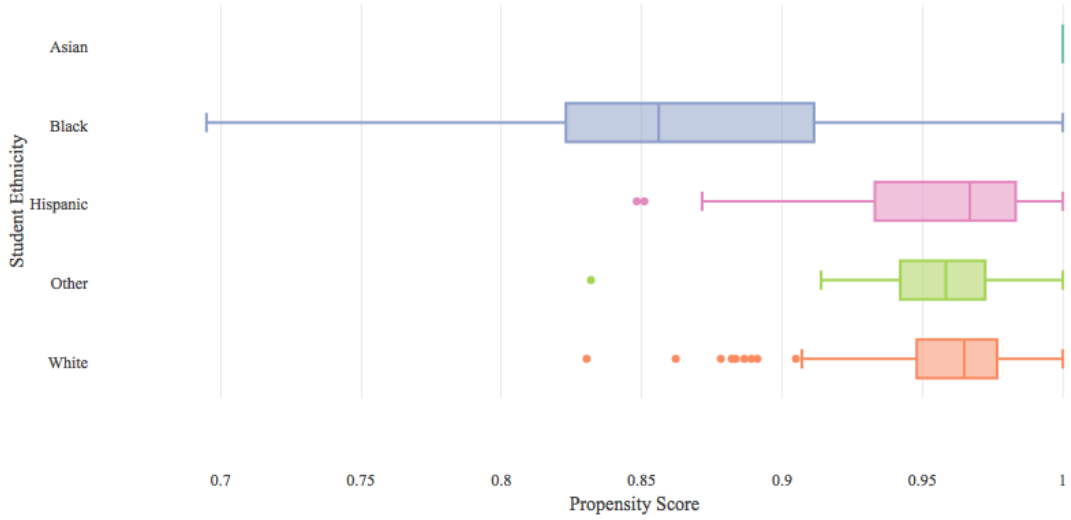
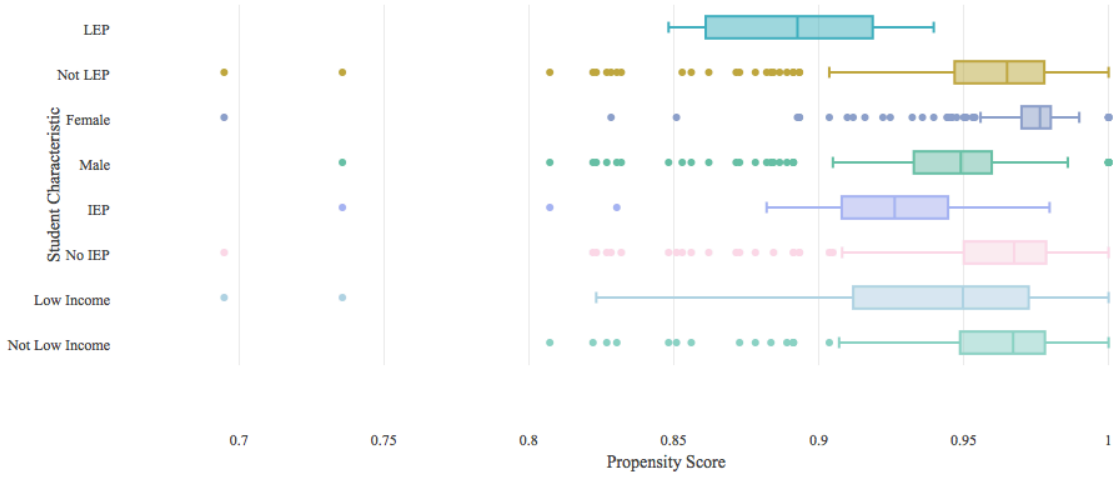


Figure 8. Propensity Scores Disaggregated by Student Characteristics



In general, LEP students were less likely to have complete student records (X^2 =26.451, p<.001), along with students with IEPs (X^2=14.049, p<.001), and students eligible for Free or Reduced Lunch (X^2=25.316, p<.001). Female students were more likely to have complete records (X^2= 9.620, p=.002). Black and Hispanic students were more likely than White students to have incomplete records (Wald=48.866, p<.001; and

Wald=12.752, $p<.001$, respectively). These differences suggest missingness occurred systematically.

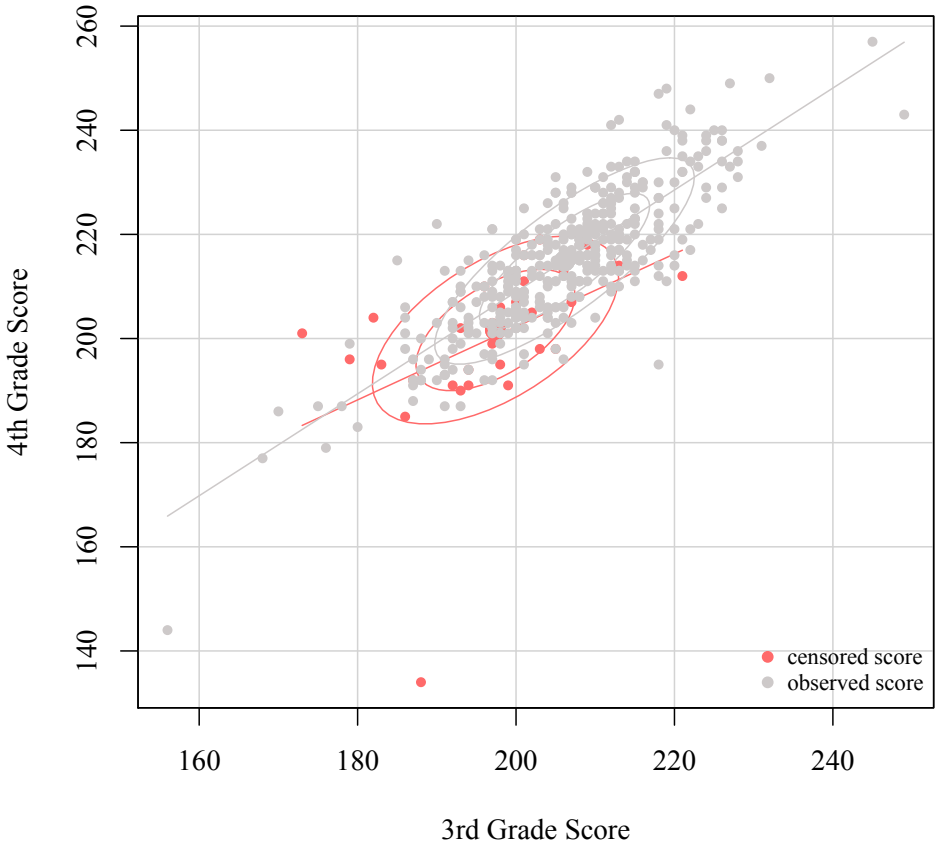
Missingness was assigned to 4th grade students in the analysis sample to mimic the patterns of missingness that were observed in the reference population. As it is likely multiple missingness mechanisms are in place simultaneously, the choice was made to assign missingness using all available student information. This design does not explicitly condition missingness on theoretical parameters determined by the researcher to fit MAR, MCAR, or MNAR assumptions. Though missing data mechanisms are unverifiable, a discussion about the plausibility of a MAR mechanism is presented in Chapter V along with implications for statistical methodology and educational policy.

Censored Sample

To determine the magnitude of missingness to impose on the complete dataset, the amount of missingness was examined in the reference population of 6th grade students. 7% of students had incomplete mathematics achievement data in the reference population; therefore 7% of current year mathematics scores were also censored in the analytic sample of 4th grade students. For the missing data methods that do not impute scores (listwise deletion and inverse probability weighting), students with censored current year mathematics scores are in effect censored at the unit level since the SGP analysis in this study uses only two variables (3rd and 4th grade mathematics scores).

Several differences were observed between censored and non-censored students. Figure 9 illustrates the relationship between the 3rd and 4th grade mathematics scores; concentration ellipses are plotted at .5 and .8 and OLS regression lines are overlaid for censored and observed student cohorts.

Figure 9. Complete Data Distributions of 3rd and 4th Grade Scores



Censored students tend to be lower in both baseline and evaluation year achievement scores, suggesting data are not missing completely at random. Missing data concentrated at the lower end of the joint distribution does not automatically imply model estimates will be biased, however it does suggest that missing data have the potential to skew estimates if not accounted for since missingness is systematic and non-ignorable.

In addition to differences in mathematics achievement data, the censored and non-censored student groups showed differences in demographic composition. Tables 5 through 9 provide frequency distributions of demographic characteristics among censored and non-censored students.

Table 5. Gender Frequencies of Censored and Non-Censored Students

Gender	Censored Status		Total
	Censored	Non-censored	
Female	7 (23.3%)	203 (52.7%)	210 (50.6%)
Male	23 (76.7%)	182 (47.3%)	205 (49.4%)
Total	30 (7.2%)	385 (92.8%)	415 (100%)

Note: Column percentages are reported.

As expected, gender differences between censored and non-censored student were observed since gender was a factor in the propensity score model used to assign artificial missingness. Female students tended to have higher propensity scores for complete data, resulting in a greater proportion of male students in the censored group.

Table 6. Free or Reduced Lunch Eligibility of Censored and Non-Censored Students

Free or Reduced Lunch Eligibility	Censored Status		Total
	Censored	Non-censored	
Eligible	15 (50%)	55 (14.3%)	70 (16.9%)
Not eligible	15 (50%)	330 (85.7%)	345 (83.1%)
Total	30 (7.2%)	385 (92.8%)	415 (100%)

Note: Column percentages are reported.

It is notable that half of FRL students were censored, despite comprising only 16.9% of the total student body. Free or Reduced Lunch eligibility is often used as a proxy for socio-economic status. As a greater percentage of censored students were FRL eligible, missing data has the potential to disproportionally impact or misrepresent low-income students.

Table 7. LEP Status of Censored and Non-Censored Students

LEP Status	Censored Status		Total
	Censored	Non-censored	
LEP	4 (13.3%)	3 (0.8%)	7 (1.7%)
Not LEP	26 (86.7%)	382 (99.2%)	408 (98.3%)
Total	30 (7.3%)	385 (92.7%)	415 (100%)

Note: Column percentages are reported.

A relatively small number of total students were Limited English Proficient (1.7%); however more LEP students were censored from the analysis sample. LEP students tended to have lower propensity scores for complete data and this manifested in in greater proportions of LEP students in the censored student group.

Table 8. IEP Status of Censored and Non-Censored Students

IEP Status	Censored Status		Total
	Censored	Non-censored	
IEP	9 (30%)	33 (8.6%)	42 (10.2%)
No IEP	21 (70%)	352 (91.4%)	373 (89.9%)
Total	30 (7.3%)	385 (92.8%)	415 (100%)

Note: Column percentages are reported.

Students with Individualized Education Plans (IEPs) comprised a greater proportion of the censored student group than the non-censored group.

Table 9. Ethnicities of Censored and Non-Censored Students

Ethnicity	Censored Status		Total
	Censored	Non-censored	
Asian	0 (0%)	8 (2.1%)	8 (1.9%)
Black	12 (40%)	5 (1.3%)	17 (4.1%)
Hispanic	8 (26.7%)	44 (11.4%)	52 (12.5%)
Other	1 (3.3%)	20 (5.2%)	21 (5%)
White	9 (30%)	308 (80%)	317 (76.4%)
Total	30 (7.2%)	385 (92.7%)	415 (100%)

Note: Column percentages are reported.

Censored and non-censored were dissimilar in ethnic composition, as non-white students were a majority of the censored group (70%) and comprised only 20% of the

non-censored group. Though demographic variables are not modeled in the SGP analysis, these characteristics are used as to pre-process missing data in each method in this study with the exception of listwise deletion. Disparities in demographic distributions are important in that they distort the representativeness of auxiliary or weighting variables.

Benchmark Analysis

The benchmark/complete data analysis serves as a benchmark comparison for the five missing data methods explored in this study. To determine how well missing data methods recover attributes of the hypothetically complete data, student growth percentiles (SGPs) derived using the complete data are compared to observed SGPs derived with each missing data method. Complete case SGPs are operationally defined to represent true or benchmark growth scores in that these are the scores that would have been observed had there been no missing data.

Figure 10. Benchmark Data: Conditional Quantile Regression Curves

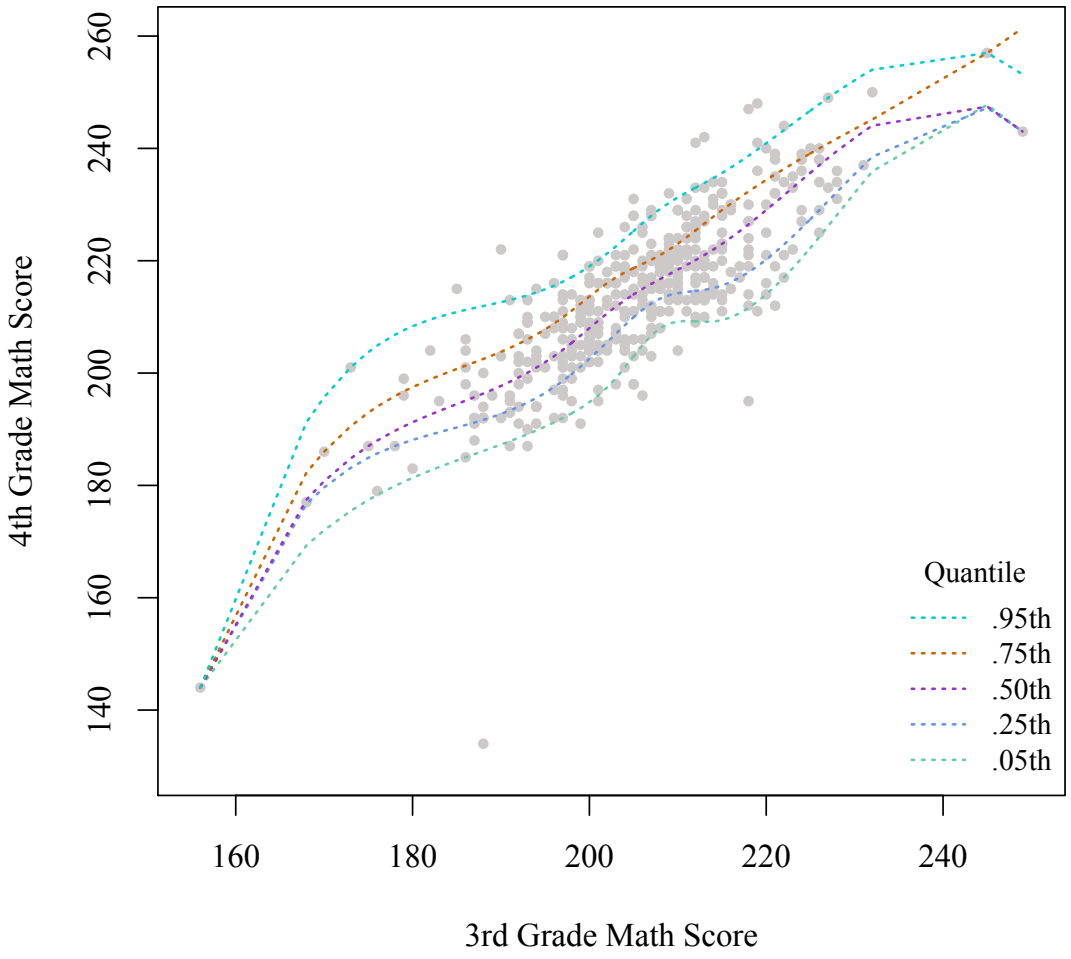


Figure 10 displays complete case conditional quantile regression estimates for the 5th, 25th, 50th, 75th and 95th student growth percentiles (τ= .05; .25; .5; .75; and .95). These quantiles are selected to visualize model estimates, however the SGP model estimates 99 quantiles total. Regression curves are particularly sensitive to extreme scores in both tails where data are sparse. There is a prominent pattern in which the quantile curves in the middle of the distribution are closer together, resembling a bottleneck. Due to this structure small differences in 4th grade scores can translate to large differences in percentile values in this part of the distribution. When data are more

compact, scores are in closer proximity to several quantile curves and thus are closer to several estimated growth percentiles. This concept is illustrated in Figure 8 (above) as the 75th and 95th quantile curves are further apart at a baseline score of 180 than they are at 200. As a result, greater gains are necessary to move from the 75th to 95th growth percentiles for students with a baseline score of 180 compared to their peers with a baseline score of 200.

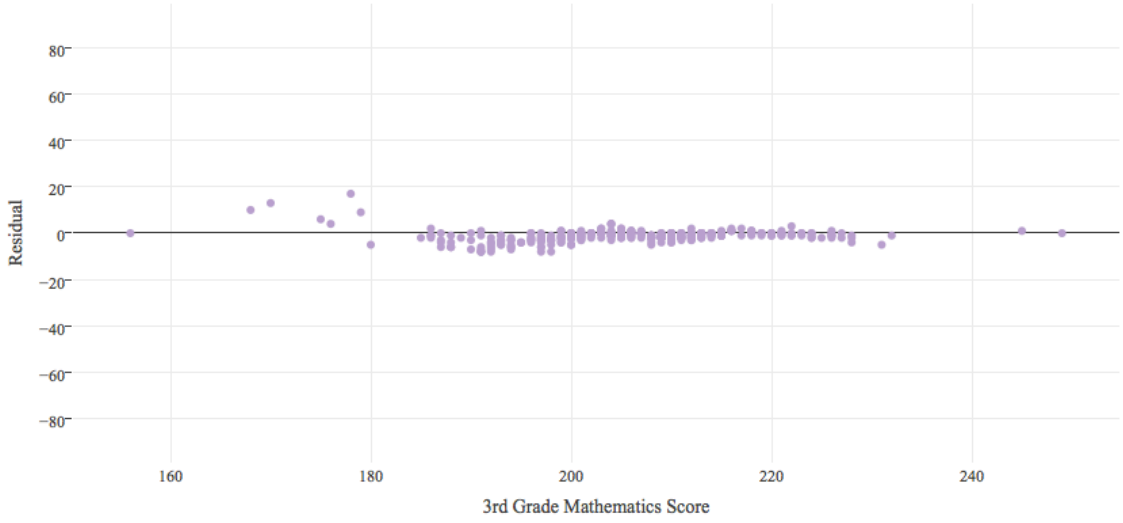
Analysis of Missing Data

Listwise deletion (LD)

The listwise deletion method ignores entire records with missing values and makes no additional adjustments for missing data before proceeding with the primary analysis of interest (in this case, the SGP model). Though missing data estimation techniques are a focus of this study, the listwise deletion method that does not estimate missing data is arguably the most important condition to investigate. This method is the default method implemented in practice.

In this study, the listwise deletion method produces SGPs for the subset of students with fully observed data only. Since a student's 3rd grade mathematics score is the single predictor variable used in the SGP model, no other academic indicators are available to help preserve attributes of the true growth distribution in the growth model. Figure 11 presents residual values produced by listwise deletion that were calculated by subtracting the true/benchmark SGP from the observed SGP.

Figure 11. Listwise Deletion: SGP Residuals by 3rd Grade Mathematics Score



No student shifted more than 19 percentile values, though missing data did cause deviations from benchmark scores visible as SGPs stray from the 0 residual line in the plot above. This dispels the notion that missing data are only an issue for teachers with missing student scores, as SGP values fluctuate among students with complete data when missing data are ignored. Negative residual values that indicate the LD model underestimated the true SGP are evident across the 3rd grade score distribution. The average residual value for the LD model was -1.19. Overall, students with higher prior achievement scores were more robust to shifting growth percentiles due to missing data as residuals for baseline scores below 200 show greater variation.

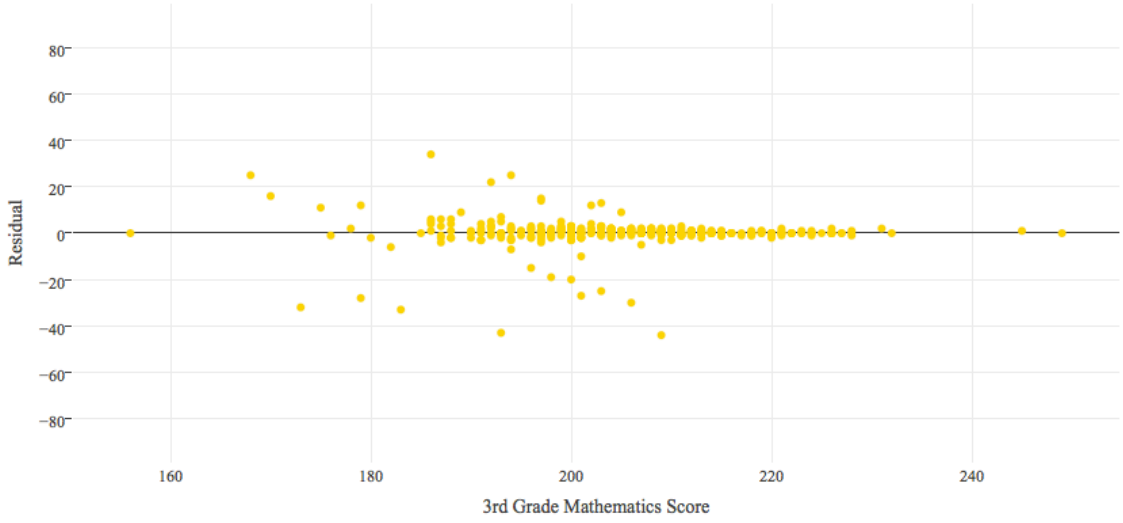
Imputation using an Expectation-Maximization (EM) Algorithm

This method takes advantage of partially complete data rather than discarding it to incrementally improve parameter estimates over several rounds of approximating missing scores. The first step in the EM imputation process is to estimate the conditional expected values for missing data using the mean vector and covariance matrix of observed data.

Observed and expected values are used to collectively update the mean vector and covariance matrix. Then, new parameter estimates are used in the next iteration to generate new expected values for missing observations, and this process repeats itself until convergence is reached. While the EM algorithm does not explicitly impute values, in this study the final parameter estimates are used to generate likely achievement scores for students with missing data given their other known information.

To arrive an imputation dataset for the SGP analysis, maximum likelihood procedures are used to estimate regression equations that predict the means, variances, and covariances with a higher accuracy than traditional regression methods. Both types of imputation models assume a multivariate normal distribution and impute values through linear regression. Similarly, both methods underestimate standard errors and require adding error to each estimate to preserve variability. Auxiliary variables including demographic characteristics and prior assessment history were used in the imputation models to "recover" the 4th Grade test scores needed to compute SGPs. Figure 12 shows SGP residuals produced by the EM imputation model across the baseline score distribution.

Figure 12. EM: SGP Residuals by 3rd Grade Mathematics Score



Noticeable in the plot above, residuals for students with higher baseline scores were smaller on average. There are several important observations to make from this plot. The EM imputation data showed a greater range in residual values than did the listwise deletion method. No student shifted more than 50 percentile values under the EM imputation method compared to 19 with listwise deletion. Unlike the listwise deletion method, residuals in this plot show less consistent bias down, as residual values bounce around the 0 residual line with both positive and negative values.

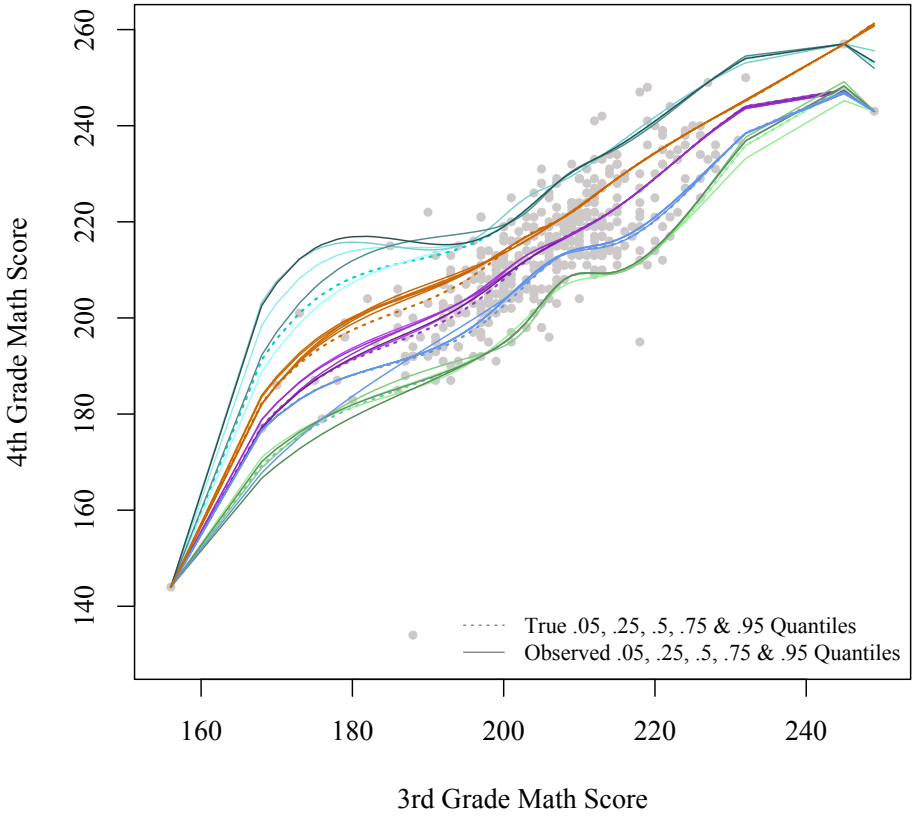
The mean residual value was .118, and the average absolute error was 2.348. Though the EM residuals show more average variability compared to the LD scenario, as baseline scores move beyond a 3rd grade score of 215, EM residuals tend to gravitate toward 0. EM model showed fairly consistent SGP estimates for higher-achieving students. Students with censored observations tended to have lower initial achievement status, where the model showed more uncertainty and less accuracy in SGP estimates.

Multiple Imputation using a Markov Chain Monte Carlo (MCMC) Method

Similar to other imputation methods, multiple imputation using a linear regression-based MCMC method utilizes incomplete data to "fill in" holes in the data to refine parameter estimates. The mean vector and covariance matrix of observed data form the prior distribution, and are used to generate initial starting values for missing data in the first iteration of the MI procedure. Imputed values are predicted using the mean and covariance matrix, and random residual error is added in to preserve variability. Alternate parameter estimates are generated using the newly-complete data, and these estimates define the posterior predictive distribution. Monte Carlo simulation draws new mean and covariance estimates from the posterior distribution generate new imputations in the next imputation step, where estimates from prior steps do not impact the current analysis (as they are "memoryless").

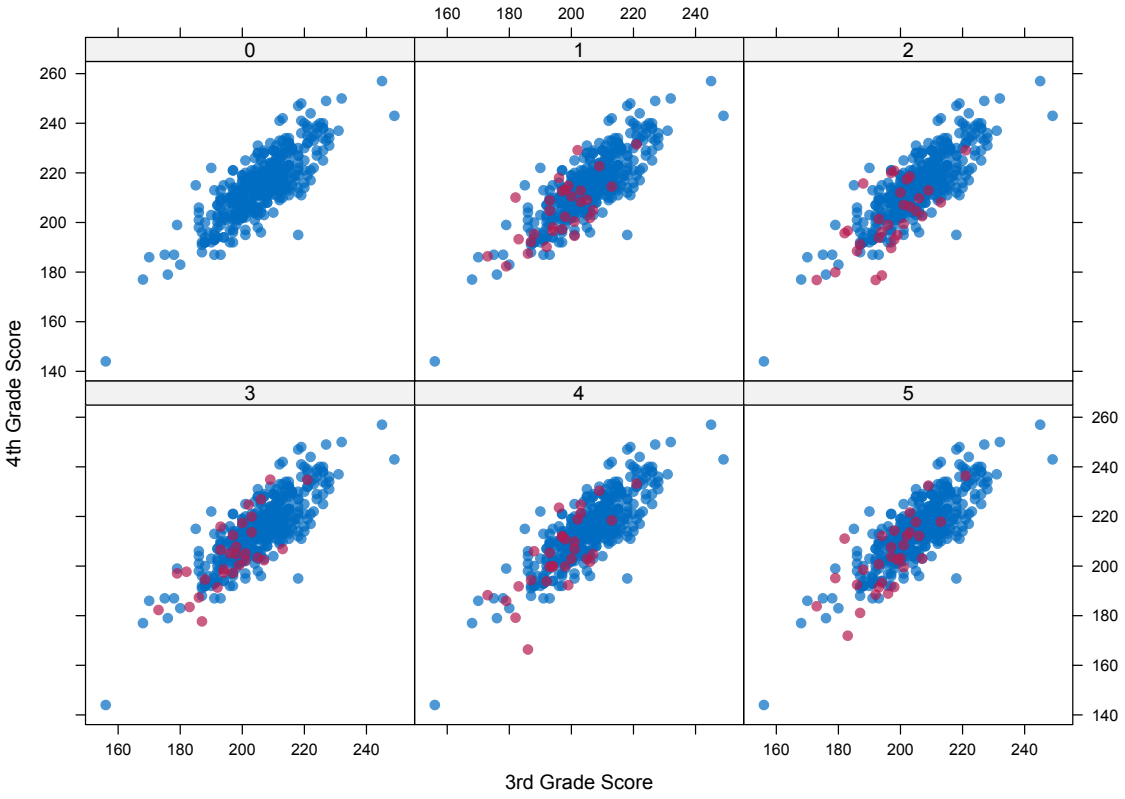
In this study, 5 imputation datasets were generated via a MCMC method and were independently analyzed through the SGP model. Resulting SGPs were converted to normal curve equivalents (NCEs) to pool imputation estimates since NCEs hold equal interval properties unlike percentiles. Pooled NCEs were then converted back to percentiles to arrive at the final growth estimates for the MCMC method. A key advantage of multiple imputation is several imputation datasets preserve variability in estimates, as displayed in Figure 13 as there are 5 curves estimated for each quantile (m=5 imputed datasets).

Figure 13. MI via MCMC: Conditional Quantile Regression Curve Estimates



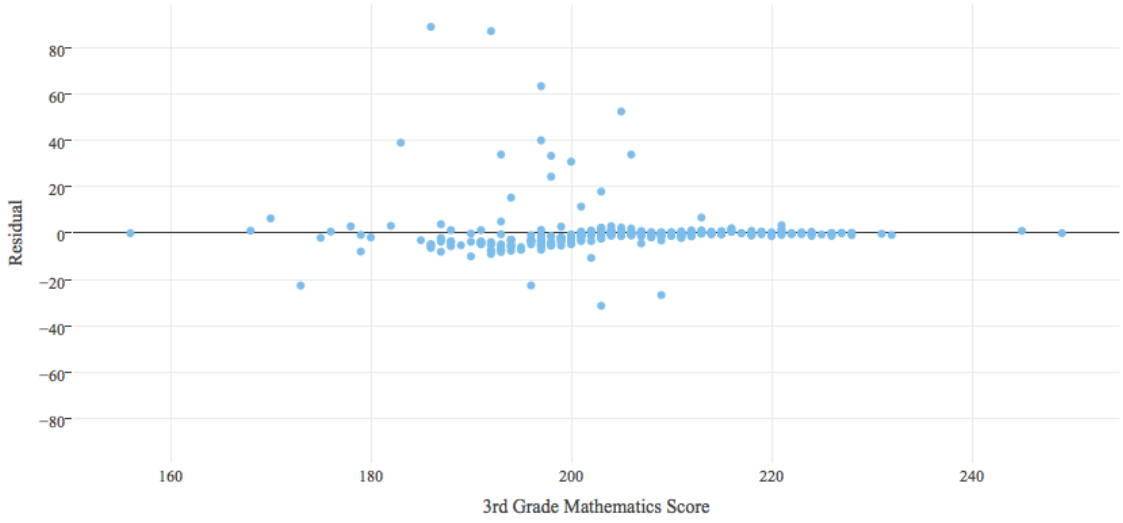
Regression splines for the 95th quantile illustrate the variability this method produces in particular. Distance between 95th percentile estimates (designated by light blue curves) is more pronounced at a baseline score of 180 than in other parts of the distribution. At a baseline score 220, curves almost overlap signifying the corresponding SGP estimates will be are similar as well. Variability in imputation estimates is also apparent in Figure 14.

Figure 14. Scatterplot of Imputed Values using a MCMC Method



Observed scores are indicated by blue dots and imputed scores are indicated by red dots. Observed data points are static whereas the imputed data change in each imputation sequence in the matrix above, reflecting the uncertainty that exists about the estimates. To evaluate the degree that these estimates reflect the complete data, residuals are plotted in Figure 15.

Figure 15. MCMC: SGP Residuals by 3rd Grade Mathematics Score



Multiple imputation estimates generated via a linear model using a MCMC method typically showed the most error at the lower end of the baseline score distribution. The mean residual SGP this method produced was 0.047.

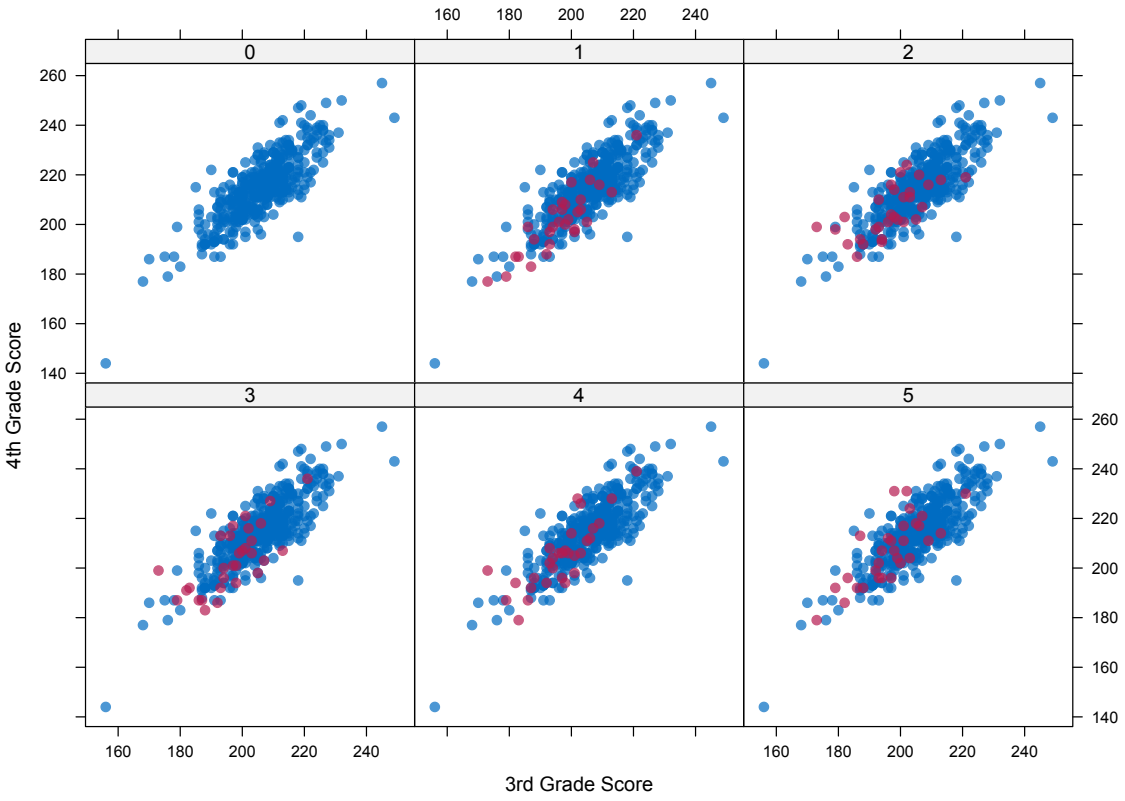
The linear relationship between 3rd and 4th grade mathematics scores is an important attribute of the study data that should be considered when evaluating the model estimates. Since linear regression is the basis of this imputation method, non-linear data may be less compatible and could impute a linear bias. The SGP analysis models curvilinear relationships between 3rd and 4th grade scores, so discordant assumptions of linearity between the imputation and analysis models could present issues with less linear data.

Multiple Imputation using a Predictive Mean Matching (PMM) Method

To generate MI estimates using a semi-parametric approach, a predictive mean matching method was implemented. A total of 5 datasets were imputed and

independently analyzed using the SGP methodology to generate 5 different growth percentile estimates for each student. Similar to the MCMC multiple imputation model, SGP estimates were converted to NCEs for pooling and were later transformed back to the percentile metric. Figure 16 shows the variability in imputation values for each imputation dataset.

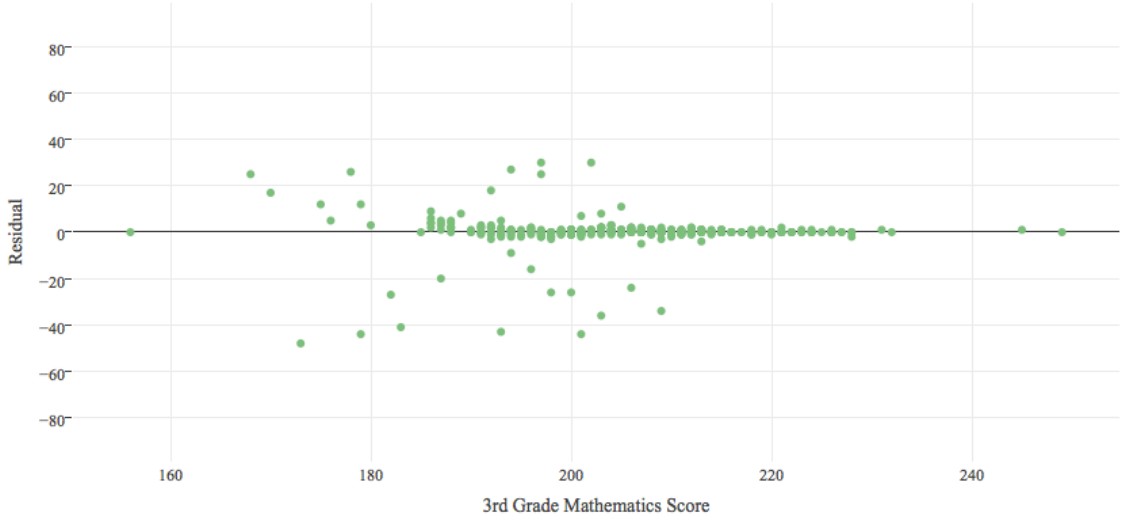
Figure 16. Scatterplot of Imputed Values using a PMM Method



Unlike estimates generated via the linear imputation model using a MCMC method, all imputations were values that originated in similar donor cases. This procedure resulted in restricted a score range (limited to that of non-censored cases) that is apparent in the plot of residual values in Figure 17. Increasing the pool of donor values to k=5 or k=10 would increase the variation in residual values, though given a small

sample, this may also result in many cases that are dissimilar to cases they are matched to, and there are no definitive guidelines to specifying a PMM model.

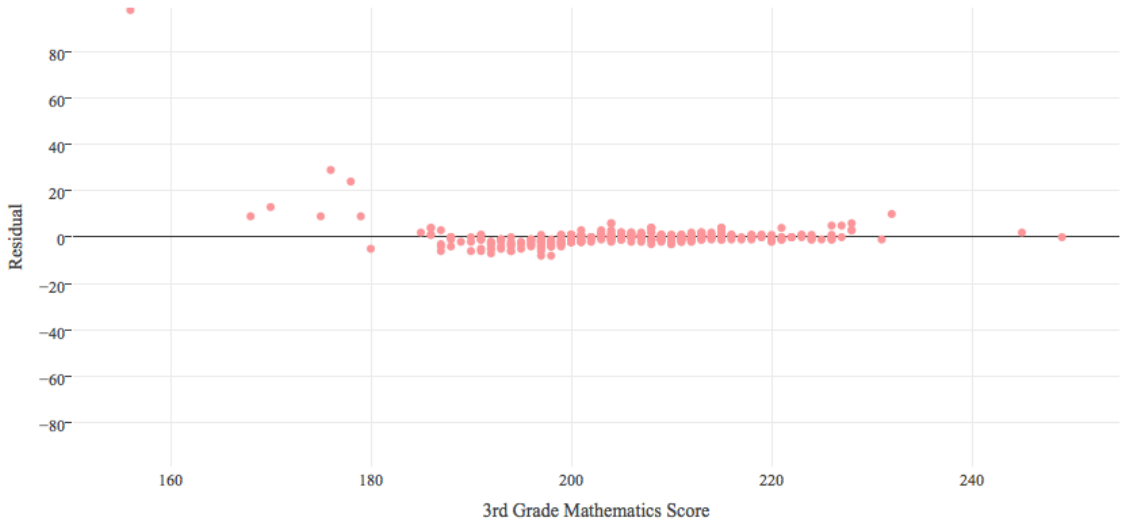
Figure 17. PMM: SGP Residuals by 3rd Grade Mathematics Score



Inverse Probability Weighting

Inverse probability is the final method posited in this study for handling missing observations. Instability of very high probabilities is a known problem of inverse probability weighting. This was not an issue in this study, as cases with high predicted probabilities of missingness were the ones that were censored from the weighted analysis.

Figure 18. IPW: SGP Residuals by 3rd Grade Mathematics Score



This model appeared to be the most similar to the listwise deletion method given the pattern of residuals. This is not surprising as both LD and IPW methods do not impute scores and only estimate SGPs for students with no missing data. In contrast, imputation methods showed less deviation from the benchmark SGP for students with higher 3rd grade achievement. One potential explanation for the dissimilarities between the IPW and imputation models is that this method is inherently different in its methodology and sample. Weighting prioritizes observations based on their likelihood of being observed in an effort to make the sample more representative of the complete data. This procedure shifts the distribution of observed values but does not attempt to fill the void left by incomplete observations, whereas imputation estimates augment the dataset and deliberately introduce stochastic error to preserve variability.

Overall Model Comparisons

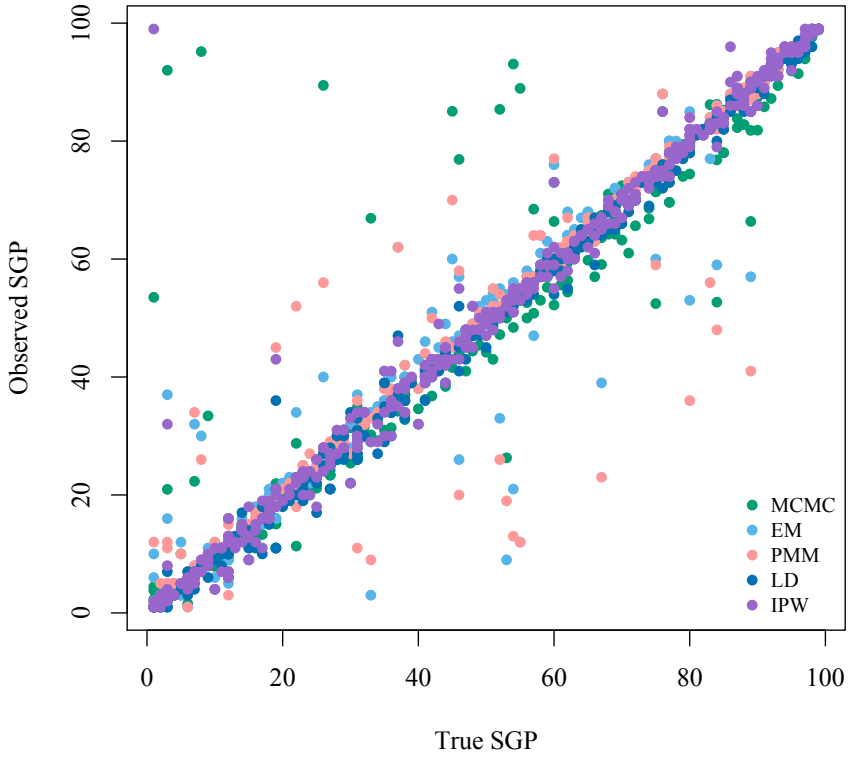
Table 10 provides correlations between student growth percentiles derived from each missing data method, and Figure 19 (below) visualizes the relationship between true and observed SGPs for each missing data method. In calculating correlations, pairwise-deletion was implemented so the listwise deletion (LD) and inverse probability weighting (IPW) correlations are included. Since LD and IPW procedures utilize only a subset of the sample, these scenarios include fewer students (n=385).

Table 10. Correlations between Missing Data and Complete/Benchmark Model SGPs

	Complete data (benchmark)	Listwise Deletion	EM	MCMC	PMM	IPW
Complete data (benchmark)		0.996***	0.978***	0.947***	0.968***	0.979***
Listwise Deletion	0.996***		0.997***	0.997***	0.997***	0.983***
EM	0.978***	0.997***		0.941***	0.992***	0.979***
MCMC	0.947***	0.997***	0.941***		0.927***	0.980***
PMM	0.968***	0.997***	0.992***	0.927***		0.981***
IPW	0.979***	0.983***	0.979***	0.980***	0.981***	

Note: Computed correlation used spearman-method with pairwise-deletion.
***p<.001

Figure 19. True and Observed SGPs by Missing Data Method



All models demonstrate a high correlation between the complete case benchmark. Listwise deletion produced the highest correlation with complete case SGPs. Since the goal of imputation and other missing data handling techniques is to preserve underlying properties of the would-be complete data as a whole, correlations calculated using the total sample of 415 students (30 partially-observed, 385 fully-observed) may not be the best indicator for overall model performance. The stochastic process of adding residual error to imputation estimates serves the purpose of preserving variability, but by design it will reduce the prediction accuracy of individual student estimates. Limiting the correlation to the 385 non-censored students increases the correlation between complete case SGPs and imputation method estimates presented in Table 11.

Table 11. SGP Correlations, Censored Observations Excluded

	Complete (benchmark)	LD	EM	MCMC	PMM	IPW
Complete (benchmark)		0.996***	0.997***	0.997***	0.996***	0.979***
LD	0.996***		0.997***	0.997***	0.997***	0.983***
EM	0.997***	0.997***		0.995***	0.998***	0.979***
MCMC	0.997***	0.997***	0.995***		0.995***	0.980***
PMM	0.996***	0.997***	0.998***	0.995***		0.981***
IPW	0.979***	0.983***	0.979***	0.980***	0.981***	

Note: Computed correlation used spearman-method with listwise-deletion.
****p<.001*

Correlation results in Table 11 show all 3 imputation models (EM, MCMC, and PMM) are more comparable to the benchmark SGP values for the non-censored cases. Imputed values were used to estimate growth quantiles of the overall distribution of 4th grade students. SGPs were not reported for students with imputed scores since predicting individual scores is not the purpose of imputation. This may represent a more useful application of imputation methods in real-world settings, as imputed scores may be misunderstood by as falsified data by parents and the larger community.

Next, the impact of each missing data method on the overall distribution of MGP estimates is explored. In the context of teacher evaluation, the median growth percentile of a teacher's students is often used as the primary summary statistic for SGP analyses. Non-parametric Wilcoxon signed rank tests (sometimes referred to as Mann-Whitney) were conducted to compare the complete/benchmark MGPs and estimates derived under each missing data method. This is a paired-sample test (analogous to the parametric equivalent of a paired-samples t-test) of the differences between benchmark MGPs and the MGPs observed under each missing data method; results are displayed in Table 12.

Table 12. Test Statistics for Observed and Benchmark MGP Differences

	LD	EM	MCMC	PMM	IPW
Z	-1.050[b]	-.338[a]	-.037[a]	-.380[b]	-.640[a]
Asymp. Sig. (2-tailed)	.294	.735	.970	.704	.522

[a]Based on negative ranks. [b]Based on positive ranks.

Differences did not reach statistical significance for any missing data method. Findings suggest that aggregate growth scores produced by each missing data method did not result in significant deviations from the benchmark/complete data values. These results are consistent with the high correlations observed between benchmark and missing data MGPs presented in Table 13.

Table 13. Correlations between Missing Data and Complete/Benchmark Model MGPs

	Complete (benchmark)	LD	EM	MCMC	PMM	IPW
Complete (benchmark)		0.900***	0.963***	0.973***	0.907***	0.918***
LD	0.900***		0.909***	0.903***	0.939***	0.994***
EM	0.963***	0.909***		0.922***	0.949***	0.930***
MCMC	0.973***	0.903***	0.922***		0.904***	0.910***
PMM	0.907***	0.939***	0.949***	0.904***		0.948***
IPW	0.918***	0.994***	0.930***	0.910***	0.948***	

Note: Computed correlation used spearman-method.
****p<.001*

Listwise deletion (LD) resulted in the lowest MGP correlation, though the PMM correlation is only slightly better by .007 compared to the benchmark data. Multiple imputation using a MCMC method showed the most similarity to benchmark MGP values, producing a correlation of .973. All methods produced a correlation of .9 or higher, showing comparable estimates of the MGP metric, through correlations using the SGP metric were slightly higher. Since MGPs are aggregated at the teacher level, one

potential explanation for somewhat deflated MGP correlations is that not all teachers were linked to students with missing data, and teachers had varying amounts of censored students. The SGP estimate is calculated at the student level and is invariant to changes in the classroom roster.

To further explore the relationship between MGPs and the frequency of censored students, rank correlation coefficients were calculated and compared. As censored student observations were not assigned randomly, the number of censored students for each teacher ranged from 0 to 4. MGPs were ranked to represent their relative standing in the overall distribution of MGP values, and then correlations were calculated between a teacher's ranking and the number of censored students in his or her class. Results of this analysis are presented in Table 14, and show relatively weak relationships between MGPs and the frequency of censored students.

Table 14. Rank Correlations between MGPs and Number of Censored Students

Rank MGP	Correlation with # of Censored Students
Complete (Benchmark)	0.00
LD	-0.11
EM	-0.22
MCMC	0.10
PMM	-0.25
IPW	-0.11

*$p<.05$

No correlations reached statistical significance setting α at the .05 level, indicating a teacher's MGP was not significantly related to the number of censored student observations linked to each teacher. Significant findings would imply estimates are biased by the frequency of missing data (e.g. teachers with higher ranked MGPs were less prone to missing student observations or vice versa).

Absolute Growth Differences

Correlations are not the only criteria for comparing growth estimates derived under each missing data method. Models with profound universal differences in observed and expected values error can still produce a high correlation. Though correlations are one measure of comparability, they do not provide information about the absolute differences in growth scores between models. The correlation metric ranges from -1 to 1 and the SGP metric ranges from 1 to 99. By itself, a correlation does not indicate how many percentile values students change (e.g. the same correlation value could represent a shift from the 1st to 2nd growth percentile values or a shift from the 1st to 52nd growth percentile).

Framing model differences using the actual SGP metric provides additional context. To supplement correlation findings, Table 15 provides the average absolute values for SGP residuals for each missing data method. Table 15 also provides the percentage of student growth scores that deviated from their corresponding benchmark SGP obtained through the complete case analysis.

Table 15. Magnitude and Frequency of Differences in SGP Estimates

	N students	Mean Absolute Residual	% students retaining Benchmark SGP
LD	385	1.764	28.8
EM	415	2.347	36.9
MCMC	415	3.199	28.0
PMM	415	2.388	47.5
IPW	385	1.948	24.4

Listwise deletion resulted in the smallest mean absolute error. Examining the frequency of SGPs that deviate from the benchmark SGP, multiple imputation via PMM produced

the greatest percentage of matching student growth scores. In the case of listwise deletion, 28.8% of SGPs matched the benchmark values despite sharing 96.4% of the same 3rd and 4th grade mathematics scores (30 of 830 scores were censored among 415 total students). Since the LD and IPW methods only produce growth percentiles for the subset of students with no missing data, Table 16 presents differences in residuals among censored and non-censored student groups so comparisons can be made using the same students.

Table 16. Differences in SGP Estimates for Censored and Non-Censored Students

		Mean Absolute Residual	% students retaining Benchmark SGP
All students (n=415)	EM	2.347	36.9
	MCMC	3.199	28.0
	PMM	2.388	47.5
Non-censored students only (n=385)	LD	1.764	28.8
	EM	1.203	39.7
	MCMC	1.558	40.3
	PMM	0.968	50.4
	IPW	1.948	24.4

Again, prediction accuracy for individual estimates is not the primary goal of the imputation process. Residual error is deliberately added to imputed values to preserve standard error. Though imputed scores show larger residuals than observed scores, there is evidence that imputation methods more accurately reflect the expected 4th grade scores for quantiles 1 to 99 of the complete data determine SGP values. Since we would expect a certain amount of residual error in imputation score estimates, separating imputed residual values from fully-observed residuals clarifies the mean absolute error (MAE) comparisons presented in Table 16 (above). Limiting the analysis to non-censored data,

inverse probability weighting and listwise deletion methods produced the highest MAE and the lowest percentages of students with matching true and observed SGPs.

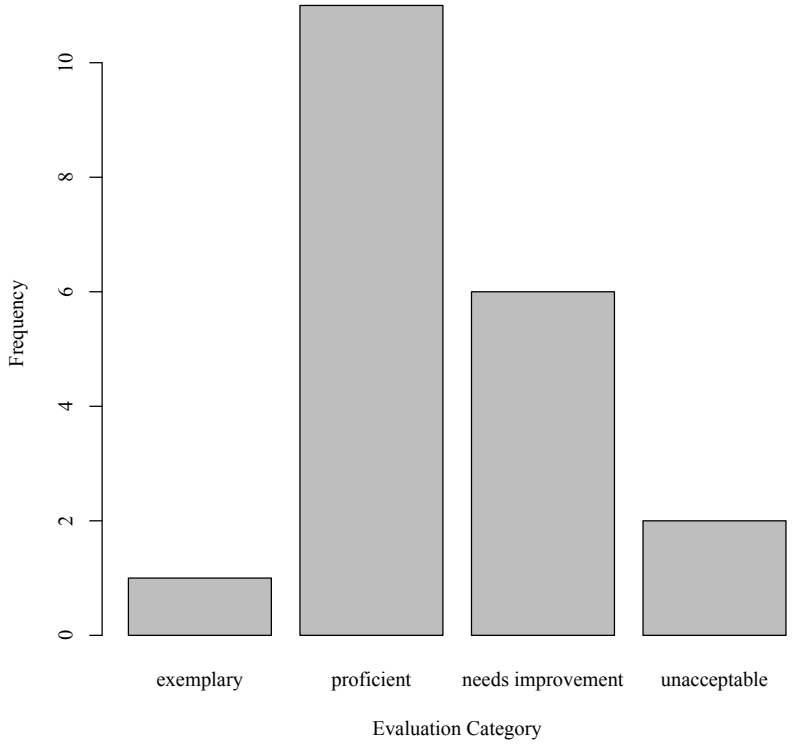
Research Question 2

Does the choice of missing data methodology result in different growth inferences when used in an educator evaluation framework?

The strong level of correlation observed between model estimates could obscure substantive differences in accountability ratings for individual teachers. Though absolute residual errors were compared to supplement correlation findings, these differences may or may not translate to different evaluation inferences when embedded in an accountability framework. Therefore, the purpose of the second research question is to document the practical implications of missing data model specification. The growth classification scheme used by the Virginia Department of Education (VDOE) was selected to demonstrate how growth scores are impacted by missing data when implemented in an evaluation context.

First it is important to consider how the evaluation scheme used in this study may impact classification rates. Similar to other frameworks, this evaluation categorization scheme is structured so that proficient ratings will be more frequent than exemplary or unacceptable ratings in most circumstances. Figure 20 shows the distribution of evaluation ratings observed in the complete case scenario.

Figure 20. Distribution of Complete/Benchmark Teacher Evaluation Ratings



In order to receive an exemplary rating, over 50% of a teacher's students would need to be classified in the highest 33% of the SGP range ("high growth" = SGPs of 66 to 99) with no more than 10% of SGPs falling within the SGP range of 1 to 34 (defined as "low growth"). Similarly, to receive an unacceptable evaluation rating, over 50% of students would need to fall in the lowest third of the SGP growth range. The proficient category requires at least 65% of student growth ratings to fall within the highest 65 SGP values ("moderate + high growth" defined as SGPs of 35 to 99). This rating scheme may reflect safeguards in place to ensure an exemplary or unacceptable classification is more difficult to obtain, analogous to giving preference to Type II error (failing to reject the

null hypothesis that growth is within the expected/mid-range when it is really low or high) versus risking the Type I error of a false positive.

Overview of Evaluation Findings

A total of 20 teachers were included in the analysis; teachers with less than 10 students in their roster were excluded. The number of censored students linked to each teacher ranged from 0 to 4. Misclassification rates for each missing data method are listed in Table 17 (calculated by row).

Table 17. Misclassification Rates

	False Positive for Unacceptable or Needs Improvement	**False Negative** for Unacceptable or Needs Improvement
LD	3 (15%)	1 (5%)
EM	1 (5 %)	0
MCMC	0	0
PMM	1 (5%)	0
IPW	1 (5%)	1 (5%)

Note: Parentheses represent proportions of the total sample.

Listwise deletion resulted in the most frequent number of misclassifications, with most biased toward a lower evaluation category. This is consistent with results from the first research aim that show the absolute residuals for the LD scenario tend to be negative (underestimating the benchmark SGP). The linear multiple imputation model (MCMC method) scenario produced the most accurate growth classifications and also had the highest correlation to benchmark MGPs.

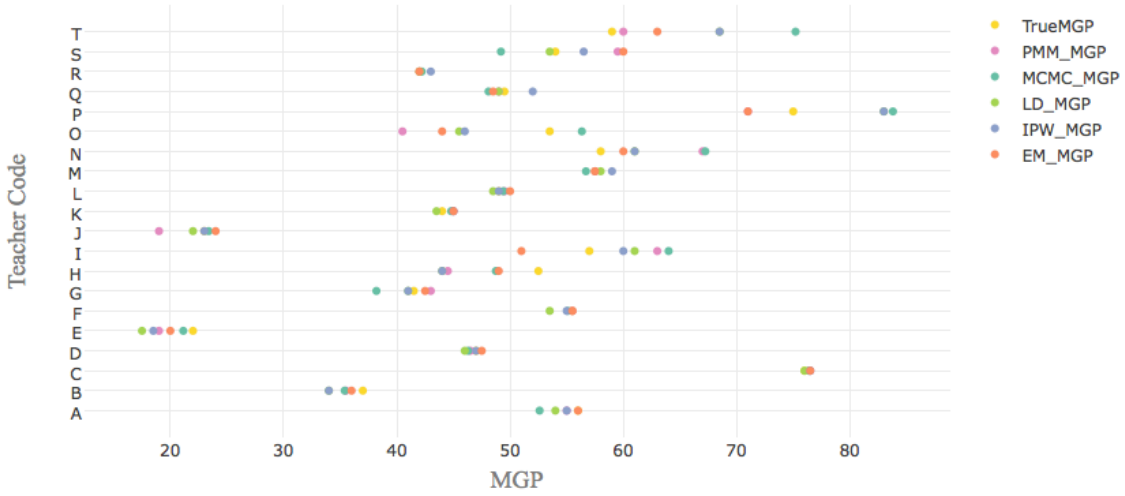
Misclassification Tolerance

Under the growth classification scheme used in this study, the MCMC estimates were robust to missing data in that no teachers resulted in a different evaluation rating as a result of model specification. The other missing data methods showed between 1 and 4

evaluation misclassifications among the 20 teachers in this study. Interpreting the practical significance of misclassification rates is less straightforward than determining statistical significance. Listwise deletion resulted in the most misclassifications, although determining whether or not this is an acceptable level requires a value judgment and is context-dependent.

Alternative evaluation frameworks use the MGP metric to define growth thresholds for evaluation. Figure 21 plots MGP estimates for each teacher across missing data scenarios.

Figure 21. Median Growth Percentile Comparisons among Teachers



Distance between coordinates along the x-axis show the dispersion of MGP estimates and could inform cut points for categorization. To consider a fixed-cut approach in categorizing growth, Table 18. presents the Massachusetts Department of Elementary & Secondary Education (MDOE) guidance for educator evaluation.

Table 18. Massachusetts DOE Growth Determinations

Evaluation Category	Growth Threshold
Low	MGP $\leq$ of 35
Moderate	35 < MGP < 65
High	MGP $\leq$ 65

This framework results in 3 misclassified teachers of the 20 teachers in this study. One teacher's MGP was misclassified by listwise deletion and IPW methods as Low; another teacher was misclassified by both multiple imputation methods as High, and one teacher was misclassified by LD, MI via MCMC, and IPW methods as High. In each of these three misclassifications, the benchmark MGP indicated Moderate growth. Using this classification approach, the only method that produced no misclassified MGPs was the imputation model using an EM algorithm.

Though the VDOE and MDOE frameworks do not indicate a superior missing data method for both contexts, they are similar in that listwise deletion was the only method to produce more than one misclassification. School systems implement plethora of different evaluation systems. Some pool MGP estimates across academic years, some weight MGPs, some set inclusion criteria for student attendance, etc., providing countless options to implement growth scores in accountability systems. The growth specification frameworks presented in this study are intended to illustrate examples of two systems and are not exhaustive.

CHAPTER FIVE

DISCUSSION

Overview

Missing student data occur for a variety of reasons, and present challenges for estimating student growth. Identifying and resolving the causes for missingness may not be realistic in practice; however several methods are available to account for missing observations when they occur. This process is especially important when student data are used in accountability frameworks and growth inferences impact evaluation decisions.

The literature on missing data methodologies for student growth models is sparse. Patterns of missingness were explored using a real dataset of mathematics achievement scores and student characteristics, and provide evidence data were not missing completely at random. This dissertation addressed two aspects of missing data in the context of student growth: 1) the comparability of missing data methods, and 2) how differences manifest when embedded in an accountability framework. These scenarios highlight the importance of both statistical and practical significance.

Summary of Findings

Research Question 1

High correlations between complete case (benchmark) growth values and estimates derived under each method act as a sensitivity analysis with respect to missing data. In general, the results favored imputation methods over deletion and weighting

approaches when the criteria are 1) the correlation to benchmark SGP values, 2) the correlation to benchmark MGP values, and 3) the smallest mean absolute error. Multiple sources of evidence suggest listwise deletion is not the best method for retaining properties of the benchmark growth distribution, though all models showed reasonably high correlations in estimates. Similarities between models are not surprising since a relatively small amount of missingness was imposed. Still, this study demonstrates the utility of missing data methods in improving growth estimates when the amount of missing observations is as small as 3.5% of achievement scores used to model student growth.

By definition, missingness is a difficult concept to measure and the reasons for missingness in any data set is largely speculated. Modeling complete and incomplete status in the reference population provides an example of how missing data manifests in the field; these findings may be useful beyond the estimation of student growth for accountability purposes. In this particular case, there is some evidence in favor of a Missing at Random mechanism rather than Missing Completely at Random since students with missing observations differed from students with complete data on some demographic characteristics.

Distinguishing a Missing Not at Random mechanism from MAR is less straightforward. MNAR models carry a different set of assumptions and require a joint model of the missingness mechanism and student growth. Misspecification of the missingness mechanism carries a different set of consequences, and it is impossible to definitely specify the cause(s) of missingness except in a simulated environment. It is possible missing student achievement data modeled in the reference population were a

function of the missing scores themselves after accounting for all other student characteristics and thus constitute MNAR. Nevertheless the missing data methods that assume MAR in this study were reasonably successful in recovering growth data.

It should also be noted that a MAR mechanism might become more plausible as additional variables are added to the model. Even if missingness can be conditioned on auxiliary variables to perfectly fulfill the assumptions of MAR, the pattern of missingness will not present as MAR if these variables are not utilized. In this study, a MAR assumption may be less plausible if fewer auxiliary variables were used to account for missing information in the imputation models. For example, not using free or reduced lunch status or other variables related to missingness in the imputation strategies may make MAR less plausible. Though each missing data method was implemented on identical datasets with the same underlying pattern of missingness, they degree to which each method supports an assumption of MAR is not the same because MAR is an assumption rather than an attribute of the data. As more is known about the nature of missing student data, we can be more confident in which variables missingness can be conditioned on and thus more confident a MAR assumption is justifiable.

To guide decisions regarding which method to use and what assumptions are supported, sensitivity analyses can be conducted to compare several approaches. If different methods produce similar findings, we can be more confident in their results. Divergent findings may highlight violated assumptions and inform the choice of missing data method moving forward or the reporting practices for student growth scores and educator evaluation ratings.

Research Question 2

Grounding methodological decisions with practical implications can inform designers of educational evaluation frameworks as they weigh tradeoffs of each method. Much of the conversation around teacher evaluation centers on the statistical methodologies that produce growth scores. Issues of reliability, measurement error, and other technical properties of student growth models are commonly addressed. However, findings related to the second research aim motivate increased attention to categorical evaluation schemes used in practice. Growth categorization can exacerbate the bias introduced by missing data. Despite the similarities reported in the first research aim, growth classifications between models were less consistent when implemented in an accountability framework.

The routine practice of implementing listwise deletion as the default method for missing data resulted in the most misclassifications for both evaluation frameworks explored in this study. Models with stricter inclusion criteria in the SGP analysis (e.g. setting minimum attendance rates) are a logical extension of this analysis to determine whether or not they result in less misclassifications. Alternatively, systematically removing students from the analysis may mirror listwise deletion, based on inclusion/exclusion criteria rather than missing data, and may bias the model in other ways. Careful consideration of each decision in estimating and categorizing growth is warranted.

Results from both research questions highlight ways missing data can impact teachers even if they are not missing student test scores in their classroom. Framing missing data discussions around bias in the overall model is important, as opposed to

limiting the discussion to cases with missing observations or teachers with missing student data. Exploring the extent to which growth estimates can shift due to missing data is important at both the individual and system levels.

Limitations and Future Research

Given the complex nature of school systems, results from this study may not generalize across different growth models, assessments, grades, magnitudes of missingness, or teacher effect sizes. As each school system is unique, it is not expected findings from this study can support automated missing data handling techniques for widespread adoption by other school systems. Instead, decisions regarding missing data must be constantly evaluated for the local context that necessitates their use. Rather than generalizing specific study findings, more broadly, this study motivates increased attention to the issue of missing data.

SGP models are most commonly implemented with much larger samples than the one used in this study, and this limitation is a threat to the generalizability of study findings. In particular, this study examined student growth and educator accountability ratings in one school, and this setting is fundamentally different from a statewide system comprised of many schools and districts. An important follow-up study is needed to explore the impact of missing data in larger samples where missingness may manifest in different ways. However, as testing opt-outs, absenteeism, student illness, and other reasons for missing data are present in school systems of all sizes, the findings from this study may generate starting points for discussion when implementing growth models that rely on incomplete student records.

As educational policy changes, the use of growth models changes in tandem. For example, growth models sometimes incorporate end of course exams and other assessments that are administered to a smaller subset of students. Future studies can investigate the impact of missing data in situations where the sample is restricted by design. Beyond the scope of this study, accountability ratings are assigned to schools and districts in addition to educators. Additional work can inform whether or not school or district effectiveness ratings fluctuate as educator ratings and student growth scores did in this study. Another logical extension of this work would be to manipulate the percent of missingness imposed on the analysis to determine if missing data methods perform differently with different amounts of available information. This information can guide practitioners as they choose a method for their specific context.

The choice was made to use real data in this study, however future simulation studies are necessary to explore the impact of missing data in different settings and circumstances. Simulation studies can isolate or manipulate certain characteristics of the data to further disentangle the impact of missing data from other attributes of student data when evaluating growth. Developing missing data methodologies and implementing them in practice are two separate procedures. Some methods may work in theoretical situations or simulations but not in practice with real, imperfect datasets. For this reason, simulation studies should be paired with evidence generated using real data.

The widespread and continued use of growth models for educator evaluation underscores the motivation for methodological research on missing data as it relates to student growth in all types of settings. Though this dissertation discusses practical and technical details of different model specifications, often these details are dissected and

debated in lieu of the philosophical rationale behind each component of VAM methodology. If paradigmatic conflict is the centerpiece of the growth discussions, future work must focus on theoretical arguments of missing data methodologies or VAMs. Even after reaching consensus on a missing data procedure from methodological and theoretical perspectives, improper implementation can undermine performance, as even the best methods can be poorly implemented.

This study emphasizes one component of value-added methodology: choice of missing data procedure. As many components collectively determine the overall validity and precision of the model, developing a general indicator of VAM fit may be useful to evaluate model adjustments (such as choice of missing data procedure). Demonstrating different missing data methodologies produce similar results that then inform similar practice decisions may give users more confidence in their implementation. On the other hand, demonstrating approaches to handling missing data lead to different results may prompt more methodological focus before making growth inferences. Sensitivity analyses using multiple missing data methods that produce either converging or diverging findings may further advance methodological research.

REFERENCE LIST

AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. (2015). *Educational researcher*. doi:10.3102/0013189x15618385

Allison, P. D. (2002). *Missing Data*: SAGE Publications.

American Institutes for Research. (2015). *2013–14 Growth Model for Educator Evaluation: Technical Report*.

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. *Alexandria, VA: Author. Retrieved from https://www. amstat. org/policy/pdfs/ASA_VAM_Statement. pdf*.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational researcher, 37*(2), 65-75.

Andrews, K. M., & Ziomek, R. L. (1998). Score Gains on Retesting with the ACT Assessment. ACT Research Report Series 98-7.

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. (2011). The Condition of Education 2011. NCES 2011-033. *National Center for Education Statistics*.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65. doi:10.2307/3701306

Ballou, D., & Springer, M. G. (2008). Achievement trade-offs and no child left behind. *Manuscript. Peabody College of Vanderbilt University. En: www. caldercenter. org*.

Betebenner, D. V., Adam; Domingue, Ben; Shang,Yi (2014). SGP: An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories.: R package version 1.2-0.0.

Betebenner, D. W. (2008). *A primer on student growth percentiles*.

Betebenner, D. W. (2009). Norm‐ and criterion‐referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42-51.

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. The National Center for the Improvement of Educational Assessment.

Briggs, D., & Domingue, B. (2011). Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the" Los Angeles Times". *National Education Policy Center*.

Brundin, J. (2014). Thousands of students protest Colorado standardized tests. Retrieved from http://www.cpr.org/news/story/thousands-students-protest-colorado-standardized-tests

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*: John Wiley & Sons.

Chen, C. (2005). *An introduction to quantile regression and the QUANTREG procedure.* Paper presented at the Proceedings of the Thirtieth Annual SAS Users Group International Conference.

Chudowsky, N., Koenig, J., & Braun, H. (2010). *Getting Value Out of Value-Added:: Report of a Workshop*: National Academies Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*: Taylor & Francis.

Colorado Department of Education. (2013). *Colorado Growth Model - Brief Report, Student Growth Percentiles and FRL Status*. Retrieved from https://www.cde.state.co.us/accountability/cgm_sgp_frl_brief

Cyr, A., & Davies, A. (2005). *Item Response Theory and Latent variable modeling for surveys with complex sampling design The case of the National Longitudinal Survey of Children and Youth in Canada.* Paper presented at the conference of the Federal Committee on Statistical Methodology, Office of Management and Budget, Arlington, VA.

Diaz-Bilello, E. K., & Briggs, D. C. (2014). Using Student Growth Percentiles for Educator Evaluations at the Teacher Level: Key Issues and Technical Considerations.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). Selecting growth measures for school and teacher evaluations. *National Center for Analysis of Longitudinal Data in Education Research (CALDAR). Working Paper, 80*.

Enders, C. K. (2012). Applied Missing Data Analysis. *Australian & New Zealand Journal of Statistics, 54*(2), 251-251. doi:10.1111/j.1467-842X.2012.00656.x

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. *Organizational Research Methods, 6*(3), 282-308.

Gabriel, T. (2010). A Celebratory Road Trip for Education Secretary. *New York Times*, A24.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*: Cambridge University Press.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28-39.

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology, 60*(1), 549-576. doi:10.1146/annurev.psych.58.110405.085530

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins, A. G. Sayer, L. M. Collins, & A. G. Sayer (Eds.), *New methods for the analysis of change.* (pp. 335-353). Washington, DC, US: American Psychological Association.

Hansen, M. G., Dan. (2015). Response to AERA statement on value-added measures: Where are the cautionary statements on alternative measures? Retrieved from http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2015/11/19-aera-value-added-measures-hansen-goldhaber

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development, 74*(5), 1368-1378. doi:10.1111/1467-8624.00612

Honaker, J., King, G., & Blackwell, M. Amelia II: A program for missing data.

Jeynes, W. H. (2007). The relationship between parental involvement and urban secondary school student academic achievement a meta-analysis. *Urban education, 42*(1), 82-110.

Jonas, D. Student Growth Percentile Model: What should we know when including student growth percentiles in a teacher's performance evaluation? : Virginia Department of Education.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*.

Koenker, R. (2005). *Quantile regression*: Cambridge university press.

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*(3), 426-450.

Landerman, L., Land, K., & Pieper, C. (1997). An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. *Sociological Methods & Research, 26*(1), 3-33. doi:10.1177/0049124197026001001

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*: ERIC.

Liang, F., Liu, C., & Carroll, R. (2011). *Advanced Markov chain Monte Carlo methods: learning from past samples* (Vol. 714): John Wiley & Sons.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the no child left behind act of 2001. *Educational Researcher, 31*(6), 3-16.

Lissitz, R. W., & Huynh, H. (2003). Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability. *Practical Assessment, Research & Evaluation, 8*(10), n10.

Los Angeles Teacher Ratings. (2010). Retrieved from: http://projects.latimes.com/value-added/

McCaffrey, D. F., & Castellano, K. E. A Review of Comparisons of Aggregated Student Growth Percentiles and Value-Added for Educator Performance Measurement.

McCaffrey, D. F., & Lockwood, J. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics, 5*(2A), 773-797.

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. 773-797. doi:10.1214/10-AOAS405

McGuinn, P. (2011). Stimulating reform: Race to the top, competitive grants and the Obama education agenda. *Educational Policy*, 0895904811425911.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology, 14*, 75-75. doi:10.1186/1471-2288-14-75

Muenchen, R. A. (2015). The Popularity of Data Analysis Software. Retrieved from http://r4stats.com/articles/popularity/

National Council on Teacher Quality. (2010). *Human capital in Boston Public Schools: Rethinking how to attract, develop, and retain effective teachers*. Retrieved from http://www.nctq.org/dmsView/Human_Capital_in_Boston_Public_Schools_NCTQ_Report

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92*(2), 263-283.

PDE, P. S. T. f. (2015). Response to PVAAS Misconceptions: District/School Reporting. Retrieved from http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PVAAS/Professional%20Development/PVAAS%20Misconceptions%20Booklet.pdf

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. (2007). *Real data analysis / edited by Shlomo S. Sawilowsky*. Charlotte, N.C: Information Age Publishing.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525-556.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel psychology, 47*(3), 537-560.

Rubin, D., Stuart, E., & Zanutto, E. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 103-116.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association, 91*(434), 473-489.

Rubin, D. B., & Wiley, I. (1987). Multiple imputation for nonresponse in surveys.

SAS Insititute Inc. (2015). SAS® EVAAS® for K-12. Retrieved from http://www.sas.com/en_us/industry/k-12-education/evaas.html

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research, 8*(1), 3-15.

Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research, 22*(3), 278-295. doi:10.1177/0962280210395740

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16*(1), 41-49.

Sherwood, B., Wang, L., & Zhou, X. H. (2013). Weighted quantile regression for analyzing health care cost data with missing covariates. *Statistics in medicine, 32*(28), 4967-4979.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj, 338*, b2393.

United States Department of Education. (2015). Every Student Succeeds Act (ESSA). Retrieved from http://www.ed.gov/essa

Wayman, J. C. (2003). *Multiple imputation for missing data: What is it and how can I use it.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.

Wilkinson, L. (1999). Statistical Methods in Psychology Journals. *American Psychologist*.

Wright, P. S. (2010). *An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education*. Retrieved from https://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Ohio-Report-Cards/Value-Added-Technical-Reports-1/An-Investigation-of-Two-Nonparametric-Regression-Models-for-Value-Added-Assessment-in-Education-S-Paul-Wright-1.pdf.aspx

Wright, S. P. (2004). *Advantages of a Multivariate Longitudinal Approach to Educational Value-Added Assessment Without Imputation*. Paper presented at the National Evaluation Institute, Colorado Springs, Colorado.

Yoon, J. (2010). *Quantile regression analysis with missing response with applications to inequality measures and data combination*. Retrieved from http://economics.ucr.edu/seminars_colloquia/2010/econometrics/Yoon%20paper%20for%2010%2025%2010%20seminar.pdf

VITA

Katherine M. Wright was born and raised outside of Houston, Texas. She completed a Bachelor of Science in Psychology at Michigan State University before matriculating at the University of Illinois at Chicago to pursue a Master of Public Health. Throughout her tenure, she was funded through a research assistantship and was awarded a Director's Scholarship. After graduating in 2010, she accepted a position at Northwestern University in the Department of Family & Community Medicine where she is currently engaged in public health and medical education research.

Katy began the doctoral program in Research Methodology at Loyola University Chicago in 2012. During her time at Loyola, she continued to work as a research coordinator and teaching assistant for the Intermediate and Advanced Biostatistics courses in the Master of Science in Clinical Investigation program at Northwestern. She authored papers in *Family Medicine* and the Centers for Disease Control and Prevention's *Preventing Chronic Disease,* and co-authored several papers that focus on education and training for individuals with psychiatric conditions. Additionally, she presented work at various national conferences including the American Public Health Association, the American Evaluation Association, and the Joint Statistical Meetings.

She currently resides in Chicago, Illinois with her boyfriend, Mark Ghesquiere, and her much-loved dog, Lizzie-bear.