THE IMPACT OF INCLUDING TEACHER AND SCHOOL CHARACTERISTICS ON PREDICTING VALUE-ADDED SCORE ESTIMATES

Lauren E. Allen

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2021

APPROVED:

Qi Chen, Major Professor
Darrell M. Hull, Committee Member
Wendy Middlemiss, Committee Member
Melissa Savage, Committee Member
Robin Henson, Chair of the Department of Educational Psychology
Randy Bomer, Dean of the College of Education
Victor Prybutok, Dean of the Toulouse Graduate School

Allen, Lauren E. *The Impact of Including Teacher and School Characteristics on Predicting Value-Added Score Estimates*. Doctor of Philosophy (Educational Psychology), May 2021, 66 pp., 10 tables, 1 figure, 1 appendix, references, 74 titles.

Value-added models (VAMs) have become widely used in evaluating teacher accountability. The use of these models for high-stakes decisions making has been very controversial due to lack of consistency in classifying teachers as high performing or low performing. There is an abundance of research on the impact of various student level covariates on teacher value-added scores; however, less is known about the impact of teacher-level and school-level covariates. This study uses hierarchical linear modeling to examine the impact of including teacher characteristics, school characteristics, and student demographics aggregated at the school level on elementary mathematics and reading teacher value-added scores. Data for this study was collected from a large school district in north Texas. This study found that across all VAMs fitted, 32% of mathematics teachers and 37% of reading teachers changed quintile ranking for their value-added score at least once across all VAMs, while 55% and 65% of schools changed their quintile ranking of value-added scores based on mathematics and reading achievement, respectively. The results show that failing to control for aggregated student demographics has a large impact on both teacher level and school level value-added scores. Policymakers and administrators using VAM estimates in high-stakes decision-making should include teacher- and school-level covariates in their VAMs.

ACKNOWLEDGEMENTS

The completion of this document marks the culmination of one of my lifelong goals. As this journey comes to an end, I look forward to new beginnings and the opened doors for unique ways to further learn, grow, and achieve.

Todd – Thank you for your patience all the long nights I spent away from home, for raising our children so I could study, and for believing I could do this.

Harper and Sam - As this aspiration becomes a reality for me, I look forward to watching you chase your own dreams. You are going to accomplish great things.

Dr. Chen – Thank you for dedication to excellence and your guidance during this process. I'm very grateful to have been able to learn under you.

To all of my former teachers – Thank you so much for all the wisdom you imparted to me. You instilled in me a love of learning that propelled me to this point.

To all of my former students - Thank you for the inspiration you provided. You are our future, and I am honored to have played a role in your education and in your life.

TABLE OF CONTENTS

Page

LIST OF TABLES AND FIGURES

Page

Tables

Figures

THE IMPACT OF INCLUDING TEACHER AND SCHOOL CHARACTERISTICS ON PREDICTING VALUE-ADDED SCORE ESTIMATES

Introduction

The Every Student Succeeds Act (ESSA) was signed into law in 2015. Building on No Child Left Behind (NCLB), its 2002 predecessor, ESSA emphasizes student and teacher accountability. Accountability goals include closing achievement gaps, English-language proficiency, and higher graduation rates. A new program, Teacher and School Leader Incentive Program, is also included in ESSA. This program offers school districts the opportunity to receive additional grants in return for incorporating a tiered or scaled pay system based on teacher quality and performance (ESSA, 2015). For many school districts, this performance-based system is implemented in the form of value-added models (VAMs). In recent years, VAMs have become widely used not only to determine performance-based pay, but also to gauge teacher effectiveness and student achievement in general (Braun, 2005; Chetty et al., 2014; Kane et al., 2008; Kane et al., 2013).

Much controversy surrounds the use of VAMs in measuring teacher accountability, pay, and contract renewal. Districts use these measures to determine teacher quality or rank among teachers, to award financial incentives, and to determine whether employment contracts are renewed. These practices have led to multiple lawsuits filed against school districts, with the most public being *Houston Federation of Teachers (Plaintiff) vs. Houston Independent School District (Defendant)* in 2015. Arguments made in this case cited the VAM used in the Houston Independent School District failed to control for student level covariates beyond the teacher's control, such as socioeconomic status and language spoken, and also violated the teacher's

14th Amendment right to due process, as teachers who were fired over the decision were not allowed to verify or challenge their value-added scores, which were estimated from the VAM using their students' scores (Amrein-Beardsley, 2019).

As part of the American Statistical Association's advocacy to improve science for policy, Morganstein and Wasserstein (2014) reviewed the use of VAMs in making high-stakes decisions regarding teacher performance appraisals and pay. After reviewing current VAM practice, VAM limitations, and consequences of making decisions based on VAM interpretation, these authors released a statement warning strongly against making high-stakes decisions based on interpretations of VAMs. In addition, there is debate in the literature regarding these models' ability to determine causation (Chetty et al., 2014; Kane et al., 2008; Kane et al., 2013; Morganstein & Wasserstein, 2014). Causal inferences made from analyses that are not capable of determining causation will lead to invalid and misleading interpretations (Keith, 2015). In this situation, these misleading inferences could cost a teacher a deduction in pay or even their job.

Not only do VAMs show a lack of consistency across models, but random error within a single sample is likely to result in teachers being misclassified as high or low performing (Schochet & Chang, 2013). Teachers misclassified as low performing may be negatively affected outside of pay or contract renewal. The reciprocal effects model states that self-concept is related to achievement and achievement is in turn related to self-concept (Marsh & Craven, 2006; Marsh & Martin, 2011). This reciprocal effect could cause an adequate teacher, who was told they were inadequate, to lose self-concept, and thereby, actually become inadequate. This would in turn negatively impact student achievement.

VAM estimates are likely to fluctuate due to a variety of factors such as student,

teacher, and school level covariates, the measures being used to gauge student achievement, and test item format (Hawley et al., 2017; Lockwood et al, 2007; Papay, 2011; Reardon et al., 2018; Schochet & Chiang, 2013). A full review of these topics is beyond the scope of this paper. For the current study, the literature review first covers a brief history of VAMs, then an investigation of factors affecting student achievement, and finally, a proposed VAM based on this review.

VAMs

In the most general sense, a VAM refers to "a statistical analysis used to measure the amount of progress students make from year to year with a district, school, or teacher" (Tennessee Value-Added Assessment System [TVAAS], 2019, p. 1). The models find a quantitative value-added score that is used to explain the additional gains one would expect a student to make by being taught by a certain teacher or attending a certain school (Raudenbush, 2004). Each individual type of VAM, such as a random effects versus fixed effects type model or simple versus layered model, has its own unique assumptions (McCaffrey et al., 2004; Raudenbush, 2004; Rose et al., 2012; Tekwe et al., 2004). For VAMs to be properly applied and analyzed, the unique assumptions for a specific model should be met (e.g., the distribution of error scores) and model specification should be correct (e.g., all appropriate covariates included; Morganstein & Wasserstein, 2014). VAMs are almost exclusively understood under the hierarchical linear model, or linear mixed model, framework (Tekwe et al., 2004). The two most widely used VAMs, the TVAAS and the Education Value-Added Assessment System (EVAAS), have used this framework (Tekwe et al., 2004; Statistical Analysis System [SAS] EVAAS, 2016). The following sections detail the origin and development of VAMs,

review the differences in models and model specification as described above, and provide an example of a generic VAM and its interpretation.

Origin and Development

The TVAAS was the first widely used VAM in the United States for the purpose of evaluating teacher, school, and school district accountability. Implemented in 1992, the TVAAS uses mixed-model equations to scale student scores on a statewide achievement exam. The TVAAS then uses these scores to compare student gains to national norms, with student and school effects considered to be fixed and teacher effects considered to be random (Sanders & Horn, 1994). The TVAAS was the first of its kind to measure student progress over time rather than static achievement in a given year, spurring a paradigm shift for educators and policymakers (TVAAS, 2019). This, along with national mandates, created the need for a more general VAM to be used nationally (TVAAS, 2019). The answer came in the form of the EVAAS. The EVAAS was modeled after the TVAAS, and both are now copyrighted by SAS. Both systems include a multivariate and univariate analysis option that can be adapted and implemented to fit individual districts' objectives (SAS EVAAS, 2016). As VAMs have become more widely used in response to ESSA and the Teacher and School Leader Incentive Program, both the TVAAS and the EVAAS have been developed and refined to be more efficient and accurate (SAS EVAAS, 2016).

Over the past 3 decades, VAMs have received considerable attention in research literature, particularly with regard to their creation, interpretation, and application. Supporters of using VAMs for teacher evaluation hold that the objectivity of evaluations using VAM scores is more effective than the subjectivity of evaluations using the school administration's yearly

observation reports (Harris, 2009). Challengers of using VAMS for teacher evaluation cite the lack of the VAMs' ability to control for confounding factors and the considerable variation in VAM scores from year to year (Hill, 2009; Koedel & Betts, 2011). To make VAMs as efficient and effective as possible, researchers have examined several areas. Multiple models have been suggested ranging from a simple single analysis to multiple, layered analyses (McCaffrey et al., 2004; Morganstein & Wasserstein, 2014; Rose et al., 2012; Tekwe et al., 2004). The impact of incorrect model specification, such as not including covariates at the student, teacher, and school level, has been studied in detail (Goldhaber et al., 2014; Palardy & Peng, 2015; Parsons et al., 2019). Specifics of this research is covered in the section to follow.

Application, Impact of Covariates, and Model Performance

Studies have been conducted to examine the impact of different aspects of model misspecification on VAMs. Student level covariates such as gender, ethnicity, socioeconomic status (often measured by free and reduced lunch [FRL]), English language learner (ELL) status, parental education level, and special education status are found to impact the teacher's value-added score (Goldhaber et al., 2014; Palardy & Peng, 2015; Parsons et al., 2019). The effect of summer, measured by comparing scores from spring to spring testing with scores from fall to spring testing, explains a substantial amount of variation in VAM estimates (Atteberry & Mangan, 2020; Palardy & Peng, 2015). Simple models that do not control for student covariates (e.g. student growth percentile models) are more likely to misclassify teachers as high or low performing than their more advanced counterparts (e.g. hierarchical linear modeling, also known as linear mixed modeling, or layered mixed modeling; Goldhaber & Theobald, 2012).

Model complexity also impacts the evaluation of teacher and school value-added scores.

The layered mixed model takes a multivariate approach allowing analyses of student scores of several subjects at once (e.g. math and reading), which then accounts for intra-student correlations, a correlation not accounted for in more simple models (Tekwe et al., 2004). Treating teacher and school effects as fixed or random also impacts the teacher and school value-added score (Raudenbush, 2004; Tekwe et al., 2004). While research shows all of these factors (i.e., complexity of the model, treatment of the effects as fixed or random, the type of covariate used and at which level, and a multilevel or multivariate method) have an impact of the evaluation on a teacher's value-added score, Tekwe et al. (2004) found the inclusion of covariates to have the largest impact on the value-added assessment of student achievement at the student, teacher, and/or school level (see also Castellano & Ho, 2013; McCaffrey et al., 2004; Rose et al., 2012).

Statistical Model for a Value-Added Model

While the Teacher and School Leader Incentive Program introduced in ESSA (2015) offered the opportunity for districts to receive additional funding in return for incorporating a scaled teacher quality-based pay system, there were not strict guidelines on how the system must look. As a result, different districts across the nation may use different models and model specifications. Additionally, the literature on VAMs covers a wide gamut of different types of VAMs and their limitations and benefits (McCaffrey et al., 2004; Raudenbush, 2004; Rose et al., 2012; Tekwe et al., 2004). Several common VAMs include simple fixed effects models, hierarchical linear models, layered mixed effects models, and student growth percentile models (Goldhaber & Theobald, 2012; Raudenbush, 2004; Tekwe et al., 2004). Simple fixed effects models are similar in nature to a single level regression analysis and use the student's change

score, or growth over time, as the outcome variable. Hierarchical linear models measure mixed effects and have anywhere from two to four levels, with students nested within teachers, teachers nested within schools, and schools nested within districts. Layered mixed effects models are multivariate in nature and also use the change score as the outcome variable. In addition, layered mixed effects models can control for student mobility by proportioning out the time a student spent in a different school. Student growth percentile models measure achievement of individual students compared to other students with similar test score histories (Goldhaber & Theobald, 2012; Raudenbush, 2004; Tekwe et al., 2004).

Even with the abundance of literature on the various models available, most researchers operate under the hierarchical linear modeling (HLM) framework (Tekwe et al., 2004). Hierarchical linear modeling analyzes and interprets data at the micro and macro levels simultaneously, honoring the ecological validity of data analysis (Bickel, 2007). Further, Rose et al. (2012) completed a methodological study comparing nine commonly used VAMs and found a three-level HLM with 1 year of pretest scores and a three-level HLM with 2 years of pre-test scores to be two of the top performing models based on their ability to recover true effects and maintain consistency.

The simplest form of a hierarchical linear model includes no predictors at any level, and is called a fully unconditional model. A model evaluating a teacher or school value-added score is likely to have three levels: students at Level 1, classes at Level 2, and schools at Level 3. This model would require three levels of equations, one each for the student level, the classroom level, and the school level (Raudenbush & Bryk, 2002). The unconditional VAM is presented as follows from Equations 1a to 1d, with Equation 1d showing the mixed model.

Level 1:	$Y_{ijk} = \Pi_{0jk} + e_{ijk}$, $e_{ijk} \sim N(0,\sigma^2)$	(1a)
Level 2:	$\Pi_{0jk} = \beta_{00k} + r_{0jk}$, $r_{0jk} \sim N(0,\tau_\pi)$	(1b)
Level 3:	$\beta_{00k} = \gamma_{000} + u_{00k}$, $u_{00k} \sim N(0,\tau_\beta)$	(1c)
Mixed Model:	$Y_{ijk} = \gamma_{000} + u_{00k.} + r_{0jk} + e_{ijk}$	(1d)

The subscripts account for the multilevel structure, with the indices *i*, *j*, and *k* denoting children, classrooms, and schools, respectively. Y_{ijk} represents the achievement of child *i* in classroom *j* in school *k*, Π_{0jk} is the average achievement of classroom *j* in school *k*, and β_{00k} is the average achievement in school *k.* e_{ijk} is the random student effect, or the deviation of an individual students score from the class average, r_{0jk} is the random classroom effect, or the deviation of classroom *jk*'s mean from the school mean, and u_{00k} is the random school effect, or the deviation of school *k*'s mean from the grand mean (Raudenbush & Bryk, 2002). When evaluating teachers' or schools' value-added scores, r_{0jk}, is teacher *j*'s value-added score and $u_{00k,}$ is school *k*'s value-added score (Palardy & Peng, 2015). As covariates are added into the model at the student, teacher, or school level, the value of these random effects (e.g., teacher and school value-added scores) will change. The change from the unconditional model to the conditional model including covariates can be calculated to find the additional percent of variance explained by adding in these predictors at their respective levels (Raudenbush & Bryk, 2002).

Since their inception, VAMs have been heavily scrutinized and researched by statisticians (Goldhaber & Theobold, 2012; McCaffrey et al., 2004; Raudenbush, 2004; Tekwe et al., 2004). Results of both empirical and methodological studies have emphasized the importance of correct model specification and appropriate interpretation. This paper focused

primarily on the importance of including student demographics as predictors at the student level, teacher characteristics at the teacher level, and both aggregated student demographics (e.g. percentage of students included in a given demographic) and school-level predictors (e.g. programs specific to schools) at the school level.

Factors Impacting Student Achievement

The list of factors that impact an individual's academic achievement is extensive. These factors come from the individuals themselves or the contexts to which these students belong, such as the classrooms (or teachers) and schools. Student level factors may be internal or linked to parental and other home factors (Bianchi & Lanclanese, 2005; Heyneman, 2005; Miley & Farmer, 2017; Waldfogel, 2012; Whiteside et al., 2017). Classroom (or teacher) and school factors may be aggregated student factors, such as percentage of students on FRL plans, characteristics unique to the teacher, or a program or design unique to the school (Hattie, 2018; Ratcliff et al., 2014; Zoda et al., 2011).

There is a general agreement in the education literature regarding the impact of a student's socioeconomic status and English language learner (ELL) status on student academic achievement (Bianchi & Lanclanese, 2005; Heyneman, 2005; Miley & Farmer, 2017; Waldfogel, 2012; Whiteside et al., 2017). In the state of Texas, where this study occurred, these two variables are reported to the Texas Education Agency (TEA) each year via the Public Education Information Management System (PEIMS) along with the state accountability test scores. The performance of these two sub-populations, students who come from homes with low socioeconomic status and students who are ELL, accounts for a portion of a school district's accountability rating in addition to the district's overall performance (TEA, 2020). These two

factors were chosen for use in this study due to their relevance in student achievement as shown in the literature and their importance in the state accountability system.

A teacher's effectiveness directly impacts student achievement. Areas such as teacher motivation level, or teacher perceptions of the school environment, may explain variation in student motivation and academic achievement (Engin, 2020; Gonzalez & Maxwell, 2018). Teacher motivation is highly correlated with teacher self-efficacy, both of which in turn impact student achievement (Engin, 2020). Gonzalez and Maxwell's (2018) study surveyed classroom teachers on their personal beliefs and observations regarding student achievement. Their findings included a general consensus among these teachers that content knowledge, self-efficacy, and classroom experience are positively associated with student achievement.

While not as commonly discussed in the literature, school characteristics such as school size or intervention programs have also been found to potentially impact student academic achievement (Hattie, 2018; Ratcliff et al., 2014; Zoda et al., 2011). Past researchers examining VAMs have suggested the impact of school characteristics should be included in future research (Basileo & Toth, 2019; Heck, 2009; Leckie, 2018; McCaffrey et al., 2004; Tekwe et al., 2004). The impact of school characteristics, teacher effectiveness factors, and the student demographic factors discussed in the previous paragraph, on student academic achievement are reviewed in the following section.

Student Demographics

Research has consistently demonstrated the impact of socioeconomic status on student achievement in high stakes testing. Since the days of NCLB, extensive research studies have found students coming from houses with lower family socioeconomic status perform lower on

high stakes testing than their peers coming from houses with higher family socioeconomic status (Bianchi & Lanclanese, 2005; Heyneman, 2005; McGown & Slate, 2019; Waldfogel, 2012). Now in the days of ESSA, the pattern continues, with research finding the difference in student academic achievement to be, on average, approximately 1.0 standard deviation lower for students from homes with lower family socioeconomic status than students from homes with higher socioeconomic status (Chmielewski & Reardon, 2016). Similarly, students for whom English is a second language perform lower, on average, than their native English speaking peers (Miley & Farmer, 2017; Waldfogel, 2012; Whiteside et al., 2017). As these students' English language proficiency increases, they are more likely to perform on par with their native English-speaking peers (Houston Independent School District, 2018, 2019).

Teacher Effectiveness

Many factors affect a teacher's overall effectiveness from their innate beliefs and personalities to their certification type. Teacher motivation level and teacher perceptions of the school environment explain variation in student motivation and academic achievement (Engin, 2020; Gonzalez & Maxwell, 2018). Similarly, teacher motivation is highly correlated with teacher self-efficacy, which in turn impacts student achievement (Engin, 2020). Past research has found teachers with traditional certifications may outperform those with alternative teaching certifications (Darling-Hammond et al, 2005). In addition, the numbers of years of experience and the degree level earned by a teacher are also predictive of student achievement, although these relationships may be very complex and not linear (Darling-Hammond et al, 2005; Irvine, 2019). For the current study, teacher certification type and teacher degree level were chosen due to their more objective nature.

Aggregated Student Level Variables

At the school level, factors affecting student achievement may be aggregated student variables. Higher poverty rates within schools are highly associated with a lower likelihood of a school meeting accountability standards on end of year high stakes testing (Hamilton et al., 2018; McGown & Slate, 2019). Baker and Johnston (2010) found even when funded equally, schools with higher proportions of students coming from homes with low socioeconomic status within a county or district performed lower than schools with lower proportions of students of such background.

School Size, Organization, and Programs

Not only do aggregated student level demographics at the school level explain variation in student achievement, but also characteristics of schools (i.e., school size, school organization, and school-wide programs) have an impact on individual student achievement (Hattie, 2018; Ratcliff et al., 2014; Zoda et al., 2011). As the enrollment numbers in an elementary school decrease, the average student performance tends to increase. This discrepancy is magnified when controlled for aggregated socioeconomic status (Zoda et al., 2011). School organization, such as traditional bell schedule versus block schedule, may also have an impact on student achievement, with students attending schools on a block schedule performing lower, on average, than students attending schools maintaining a traditionally scheduled day (Ratcliff et al., 2014). The implementation of school-wide intervention programs, such as free half- or full-day pre-K programs provided by the district or Head Start programs, which support children ages birth to 5 years old in low-income families and is provided by the U.S. government (Office of Head Start, 2020), may also positively impact student achievement (Hattie, 2018). Hattie's

(2018) meta-analysis of influences on student achievement found school music programs and school bilingual programs as likely to have a positive impact on student achievement. These two types of programs were examined in this study.

The Current Study

Proposed VAM

Many researchers have studied the effects of student level covariates, such as student socioeconomic resources, minority status, and the education of the parents, on teacher value added scores (Goldhaber et al., 2014; Heck, 2009; Leckie, 2018; Palardy & Peng, 2015; Parsons et al., 2019). Teacher and school level covariates have also been studied, but typically the research consists of aggregated student level data only (e.g. the percentage of students in a class or school on FRL, the percentage of parents with various educational backgrounds, or percentage of students of different ethnicities) rather than true teacher or school level covariates (Goldhaber et al., 2014; Heck, 2009). Much of the current literature has included student related covariates only; however, several authors have suggested teacher and school characteristics should be included in future research (Basileo & Toth, 2019; Heck, 2009; Leckie, 2018; McCaffrey et al., 2004; Tekwe et al., 2004). As with previous research, this study included aggregated student level data as covariates at the school level, but added to the literature by including teacher and school characteristics as covariates at the teacher and school level, respectively.

Research Questions

Through the present study, we examined the impact of including teacher and school

level covariates on teachers' value-added scores based on gains in student reading and math achievement test scores. Teacher and school value-added scores derived from student academic achievement gains when taught by a teacher having a specific certification type or level of college degree, attending a school offering a specialized learning program, or attending a school with certain school level demographics are compared. Comparisons are based on correlations between the teacher value-added scores from each model fitted in the analysis section and the change in the rankings of the teachers and schools based on these teacher value-added and school value-added score estimates from each model fitted in the analysis section. The research questions for the current study were:

1. What is the impact of failing to include teacher characteristics on elementary mathematics and reading teachers' value-added scores?
2. What is the impact of failing to include school characteristics on elementary teacher' value-added scores?
3. What is the impact of failing to include school characteristics on elementary schools' value-added scores?

To answer these questions, five hierarchical linear models were fit in the sequential process described in the Analysis Overview section for each academic outcome, and comparisons were made between the models.

Methods

Participants

Participants in the study were sampled from a large school district in an urban area of north Texas. The sample included 4,626 fourth through sixth graders studying under 103 math teachers and 4,514 fourth through sixth graders studying under 116 reading teachers during the

2018–2019 school year. All students and teachers were attending or employed by, respectively, 20 elementary schools. The composition of the sample of students studying under the mathematics teachers is .2% American Indian, 6.8% Asian, 18.4% African American, 32% Hispanic, 2.4% Pacific Islander, 34.4% White, and 5.8% two or more races. The composition of the sample of students studying under the reading teachers is .2% American Indian, 6.9% Asian, 18.4% African American, 32.1% Hispanic, 2.5% Pacific Islander, 34.2% White, and 5.7% two or more races. In the sample, 57.9% of students studying under the math teachers and 58% of the students studying under the reading teachers come from homes with lower family socioeconomic status. In the sample, 19.5% of students studying under the math teachers and 19.7% of the students studying under the reading teachers are learning English as a second language. These students were selected from a population of 5,518 fourth through sixth graders during the 2018-2019 school year. Table 1 displays the demographics of the student sample.

Table 1

Demographic Variables of Student Sample

		Mathematics	Reading
Ethnicity	American Indian/Alaskan Native	9(.19)	9(.20)
	Asian	315(6.81)	311(6.89)
	Black/African American	853 (18.44)	830(18.39)
	Hispanic	1480(31.99)	1450(32.12)
	Native Hawaiian/Pacific Islander	113(2.44)	111(2.46)
	Two or More Races	267(5.77)	258(5.72)
	White	1589(34.35)	1545(34.23)
Male		2384(51.53)	2324(51.48)

(table continues)

		Mathematics	Reading
Female		2242(48.47)	2190(48.52)
LEP		904(19.54)	890(19.72)
FRL		2676(57.85)	2617(57.98)
Grade Level	4th Grade	1559(33.70)	1502(33.27)
	5th Grade	1544(33.38)	1518(33.63)
	6th Grade	1523(32.92)	1494(33.10)
Total number of students		4626	4514

Note. Entries are in the *n* (%) format

Data Collection

The state of Texas requires all students in Grades 3–8 to take a set of standardized assessments called the State of Texas Assessments of Academic Readiness (STAAR) every spring. Mathematics and reading are assessed in the spring of each of these years, with writing added in during spring of fourth and seventh grade, science during the spring of fifth and eighth grade, and social studies during the spring of eighth grade (TEA, 2019). Mathematics and reading are the only two subjects consistent across all years, and were chosen in this study for that reason. The data from these tests are collected by the school districts each year and reported to the state for accountability evaluation. Student scores, student demographics, test administrator, and school are all detailed in the report. At the elementary level, the test administrator for a given student is typically his or her teacher of record for that subject; however, there are exceptions. The test administrator listed in the state reports was compared and matched to teacher of record listed in classroom rosters to ensure accuracy.

Teacher level data was retrieved through an open records request to the TEA. The TEA holds teacher certification information and teacher education information as public record. Each teacher's certification type, alternative or standard, and highest degree earned (i.e.,

bachelors, masters, or doctorate) was matched to the teacher of record information from the STAAR data described above. Name changes (e.g., due to marriage) were cross referenced with school records to ensure accuracy.

School level data was retrieved using two separate methods. Student demographics from the data collection process described above were aggregated using SPSS (IBM, 2019) to create school-level student demographic covariates. School characteristics data was retrieved via the school district's website and recorded into the data file.

Measures

STAAR

The STAAR test in each subject area is based on a set of curriculum standards unique to Texas, known as the Texas Essential Knowledge and Skills (TEKS). STAAR tests in each subject are administered in the spring of every school year. Mathematics and reading are the only two subjects measured consistently across the STAAR tested grade levels. While math is consistent from third to eighth grade, the different mathematics courses taken for advanced and on-level students in Grades 7 and 8 begins to vary greatly. For this study, the sixth grade STAAR data was the latest grade level chosen. Third grade STAAR data was not used as those students would have been in second grade the previous school year and, therefore, not tested with STAAR assessments. This study used each student's 2018–2019 mathematics and reading STAAR scores for fourth, fifth, or sixth grade as the outcome variable and 2017–2018 mathematics and reading STAAR scores for third, fourth, or fifth grade as a predictor variable. The coefficient alphas for the 2017–2018 reading and mathematics tests were .89 across all grades. The coefficients alphas for 2018–2019 reading tests were .89, .89, and .90 for fourth,

fifth, and sixth grades, respectively, and the coefficient alphas for the 2018–2019 mathematic tests were .90, .91, and .90 for fourth, fifth, and sixth grades, respectively.

In order for the measurement of students' growth, or change in academic achievement, from year to year to be meaningful, it is necessary for the 2018–2019 mathematics and reading test scores to be on the same scale with the 2017–2018 mathematics and reading test scores. If tests are not created with a vertical scale, a scaling process must be completed, such as creating normal curve equivalencies as done in all TVAAS and EVAAS (SAS EVAAS, 2016). The TEA (2013) created all STAAR exams on a "vertical scale score system that allows for direct comparison of student test scores across grade levels within a content area" (p. 3), and no scaling was needed in the current study.

Socioeconomic Status

Socioeconomic status was measured by a student's FRL eligibility, which was labeled dichotomously as 1 or 0 for either eligible for FRL or not eligible, respectively.

English-Language Proficiency

English-language proficiency was measured by inclusion in the district's limited-English proficiency (LEP) intervention program. This was labeled dichotomously as 1 or 0 for either a participant in the LEP program or not a participant, respectively.

Certification Type

The teacher's certification type is either "alternative", indicating they did not complete a traditional university certification program, or "standard", indicating they did complete a

traditional university certification program. This was labeled dichotomously, 1 for traditional and 0 for alternative certification.

Degree Level

The teacher's highest earned degree as of the 2018–2019 school year, as indicated in the TEA records. This was labeled dichotomously, 1 for a master's degree and 0 for bachelor's degree.

School of Choice Program

The district in this study offers School of Choice (SOC) programs at 12 of its 20 elementary schools. The district in the study defined their SOC programs as programs built to enable students to focus on special, specific topics that will grow their artistic talents and develop skills needed to compete with a global workforce. The two SOC programs offered at the elementary level include a Spanish immersion program and a Suzuki strings program. The SOC variable used contrast codes to analyze the impact of the two programs compared to each other in addition to analyzing the impact of a school offering a SOC program or not. Table 2 gives the coding scheme.

Table 2

Contrast Coding Scheme for SOC Variable

School of Choice Program	C1: ANYSOC	C2: COMPARESOC
Spanish Immersion	1/3	1/2
Suzuki Strings	1/3	-1/2
No School of Choice	-2/3	0

- *Spanish Immersion*: Participation in the Spanish immersion program begins in first

grade. Students accepted into the program are taught by a native-Spanish speaking teacher who covers all core content (i.e., math, reading, social studies, and science) speaking only in Spanish. Students in this program attend their music, art, and physical education classes with the rest of the first-grade population and English is spoken at this time. To be accepted into the program, parents enter their child into the lottery system in January of the child's kindergarten year and students are randomly chosen from this pool of entries until all seats are filled. Because of the complete immersion in a different language, the district has indicated finding a pattern of students falling behind in English compared to their peers in traditional classes the first few years of the program. Per district policy, parents entering their children into the program must sign a waiver stating their responsibility to keep their child fluent in English, including reading in English 20 minutes per night.

- *Suzuki Strings*: The Suzuki strings program allows students the opportunity to have formal training on the violin, viola, cello, or string bass. The program requires the student to attend one private lesson each week during the school day and one group session each week after school. Parents must supply the student's instrument, musical books, and any other necessary supplies. In addition, a parent is required to attend the weekly private lesson with the student. Participation in this program also requires parents to enter their children into a lottery system, from which entries are drawn until the classes are all filled.

Aggregated Socioeconomic Status and English-Language Proficiency

The percentage of students eligible for FRL at each school was used as a school-level covariate. The percentage of students enrolled in the district's LEP program at each school was also used as a school-level covariate. Both of these variables were grand mean centered at the

teacher and school level, making the 0 represent the average amount of students receiving FRL or enrolled in LEP.

Analysis Overview

The same set of data analysis was completed for mathematics and reading separately. The data analysis began with conducting descriptive statistics for the sample, followed by a model building process. All models described in the following model building process were fit using full maximum likelihood estimation, as recommended by Raudenbush and Bryk (2002) for three level hierarchical linear models.

Model Building

The model building process used in this study was designed following the work of Palardy and Peng (2015). The model building consisted of a base model (Model 1), followed by four subsequent models, each including the covariates of the previous, plus a new set of covariates: a student demographics model (Model 2), a teacher characteristics model (Model 3), an aggregated student demographics model (Model 4), and lastly, a school characteristics model (Model 5). Each model is formatted following the three-level hierarchical linear model as shown in Equations 1a through 1d, but with covariates added as follows (see Table 3): First, Model 2 was fit including only the student's prior year mathematics and reading STAAR score as a covariate. Second, Model 2 that included student demographic covariates at Level 1 was fit. Third, Model 3 that added in teacher characteristics at Level 2 was fit. Fourth, Model 4 that added in aggregated student demographics at Level 3 was fit. Finally, Model 5 that added in school characteristics at Level 3 was fit.

The change in the percentage of the variance explained in progression of the model building was also calculated. The results from the five models were compared using Spearman's rank correlations for all possible pairs of correlations of both the estimated teacher value-added and school value-added scores from each model. After this, the value-added scores were divided into quintiles at the teacher and school level, and the percentage of changes across quintile levels for the teacher–level and school-level from each model to the subsequent model was compared. Finally, the percent of overall changes among teachers and schools was calculated (e.g. if a teacher's overall movement was between one, two, or more quintiles).

Table 3

Model Building

Models	Level 1 Covariates	Level 2 Covariates	Level 3 Covariates
Base Model	Prior achievement		
Student Demographic Model	Prior achievement, FRL, LEP		
Teacher Characteristic Model	Prior Achievement, FRL, LEP	Certification Type, Degree Level	
Aggregated Student Demographics Model	Prior Achievement, FRL, LEP	Certification Type, Degree Level	Percent students FRL, Percent LEP students
School Characteristics Model	Prior Achievement, FRL, LEP	Certification Type, Degree Level	Percent students FRL, Percent LEP students, Contrasts 1 and 2

Note. Contrasts 1 and 2 are contrast coded variables as described in Table 2.

To determine if slope coefficients should be fixed or random, we calculated Akaike information criterion (AIC), Bayesian information criterion (BIC), consistent AIC (CAIC), and sample-size adjusted BIC (SABIC) fit indices for all models. Because these represent the degree of inaccuracy, or "badness of fit," in the model, the model with the lowest number for these

indices is considered to be the best model (McCoach & Black, 2008). Across all models, the lowest fit index occurred when the slope coefficients were fixed. For this reason, in all model specifications, intercept coefficients are set as random, and slope coefficients are set as fixed.

Model Specification

The multilevel equations for the final school characteristics VAM is presented as follows from Equations 2a through 2c. All of the other three models are reduced forms of this model, with different sets of covariates included as shown in Table 3.

Level 1	$Achieve_{ijk} = \Pi_{0jk} + \Pi_{1jk}*(LEP_{ijk}) + \Pi_{2jk}*(FRL_{ijk}) + \Pi_{3jk}*(PriorAchieve_{ijk}) + e_{ijk}$	(2a)
Level 2	$\Pi_{0jk} = \beta_{00k} + \beta_{01k}*(Cert_{jk}) + \beta_{02k}*(Degree_{jk}) + r_{0jk}$ $\Pi_{1jk} = \beta_{10k} + \beta_{11k}*(Cert_{jk}) + \beta_{12k}*(Degree_{jk})$ $\Pi_{2jk} = \beta_{20k} + \beta_{21k}*(Cert_{jk}) + \beta_{22k}*(Degree_{jk})$ $\Pi_{3jk} = \beta_{30k} + \beta_{31k}*(Cert_{jk}) + \beta_{32k}*(Degree_{jk})$	(2b)
Level 3	$\beta_{00k} = \gamma_{000} + \gamma_{001}*(pctFRL_k) + \gamma_{002}*(pctLEP_k) + \gamma_{003}*(C1_k) + \gamma_{004}*(C2_k) + u_{00k}$ $\beta_{10k} = \gamma_{100} + \gamma_{101}*(pctFRL_k) + \gamma_{102}*(pctLEP_k) + \gamma_{103}*(C1_k) + \gamma_{104}*(C2_k)$ $\beta_{20k} = \gamma_{200} + \gamma_{201}*(pctFRL_k + \gamma_{202}*(pctLEP_k) + \gamma_{203}*(C1_k) + \gamma_{204}*(C2_k)$ $\beta_{30k} = \gamma_{300} + \gamma_{301}*(pctFRL_k) + \gamma_{302}*(pctLEP_k) + \gamma_{303}*(C1_k) + \gamma_{304}*(C2_k)$	(2c)

The Level 1 outcome variable, $Achieve_{ijk,}$ is student *i*'s 2019 STAAR score in mathematics or reading. The model controled for the 2018 STAAR score in either reading or mathematics ($PAchieve_{ijk}$), the FRL status (FRL_{ijk}), and the LEP status (LEP_{ijk}). All slope coefficients were fixed. Π_{0jk} represented the conditional mean of the outcome for all students taught by teacher *j* in school *k*, given they scored at the average percentage point last year (indicated by a normal curve equivalent zero 2018 score), did not receive FRL, and were not enrolled in the district LEP program. Π_{1jk} was the expected change in 2019 STAAR score as the 2018 STAAR score increases by one for students not receiving FRL and not in LEP. Π_{2jk} was the difference in the 2019 STAAR score for non-LEP students receiving FRL compared to non-LEP students not receiving FRL and

who scored at the average percentage point in 2018 (i.e.,0). Π_{3jk} was the difference in the 2019 STAAR score for non-FRL students enrolled in the district LEP program compared to non-FRL students not in the district LEP program and who scored at the average percentage point in 2018 STAAR (i.e., 0). e_{ijk} was the random student effect that represents the deviation of student *ijk*'s score from the predicted score based on the student level model. e_{ijk} was assumed to be normally distributed with a mean of zero and a variance of σ^2. This analysis focused mainly on the factors contributing to the intercept Π_{0jk}, so the following explanations of the Level 2 and Level 3 models focus on the first equation in each level.

In the Level 2 model, Π_{0jk} was the conditional mean of the outcome for all students taught by teacher *j*. All slope coefficients remain fixed. β_{00k} represented the predicted 2019 STAAR scores for each *k* school with classroom having an average percentage of students receiving FRL and in LEP (i.e., 0). β_{01k} and β_{02k} represented the expected change in Π_{0jk} as the percentage of students in the classroom who receive FRL or are enrolled in the LEP program, respectively, increased by one percentage point holding the other variable constant. r_{0jk} represented the random teacher effect, or the deviation for each *j* teacher from the conditional mean achievement of the *k* school. This teacher effect, r_{0jk}, was the corresponding teacher's value-added score and the focus of this analysis. The random effects at Level 2 were assumed to have a multivariate normal distribution$\left[r_{0jk}\right]\sim \mathrm{N}\ \{[0], \tau_{\pi 00}\}$.

In the Level 3 model, all slope coefficients remained fixed. γ_{000} was the unweighted mean of the three groups (i.e., schools offering Spanish immersion, schools offering Suzuki strings, and schools offering no SOC program) tested by the contrasts, when the schools have an average STAAR score (i.e., 0), average percentage of students receiving FRL, and average

percentage of students in LEP. $C1_k$ tested the hypothesis that there are differences in STAAR scores for schools offering versus those not offering a SOC program, and the coefficient γ_{003} represented the average difference in the 2019 STAAR scores between these two types of schools, holding other variables constant. $C2_k$ tested the hypothesis that there are differences between schools offering the two SOC programs, and the coefficient γ_{004} represented the difference in the 2019 STAAR score for students attending schools offering a Spanish immersion program compared to students attending schools offering a Suzuki strings program, holding the other variables constant. γ_{001} represented the expected change in the 2019 STAAR score as the percentage of students in school *k* receiving FRL increased by one percentage point, holding the other variables constant. γ_{004} represented the expected change in the 2019 STAAR score as the percentage of students in school *k* enrolled in LEP increased by one percentage point, holding the other variables constant. u_{00k} represented the random school effect, or the deviation for each *k* school from the conditional unweighted mean of the three types of schools. This school effect, u_{00k}, was the corresponding school's value-added score and was worth examining. The random effects at Level 3 are assumed to have a multivariate normal distribution$[u_{00k}] \sim \mathrm{N}\{[0], \tau_{\beta 00}\}$. It was further assumed that σ^2, τ_π, and τ_β are independent of each other.

OLS vs. EB Residuals as Value-Added Scores

There are arguments in the literature for using ordinary least squares (OLS) residuals or empircal Bayes (EB) residuals as the teacher or school value-added scores. Several authors, including SAS in the EVAAS calculations, used EB residuals as their estimates (Castellano & Ho, 2013; Goldhaber et al., 2014; SAS EVAAS, 2016). Leckie (2018) confirmed EB residuals are

traditionally used, but cautioned these residuals may be biased due to regression towards the mean. Similarly, Schochet and Chiang (2013) found EB estimates are more likely to return both a false positive (i.e., indicating a teacher is effective when they are not) and a false negative (i.e., indicating a teacher is not effective when they really are) than OLS estimates. Because the majority of VAM research, and moreover, the majority of VAM applications, use EB residuals, this study also used these residuals, but it was worth noting these precautions.

Results

To display results, the first section evaluates the descriptive statistics and correlations among predictor variables. Second, the fixed and random effects, variance explained, and model fit for the mathematics and reading achievement models are evaluated. Third, correlations between the five models are examined. Finally, the teacher and school value-added scores are examined through both Spearman rho correlations and comparison of change in ranking across quintiles. Because the teacher and school value-added scores are the teacher-level residuals and the school-level residuals, respectively, across the five models, the words "value-added" and "residual" are often used interchangeably. "Teacher" and "classroom" are also used interchangeably.

Descriptive Statistics and Correlations

Table 4 gives the correlations and descriptive statistics for the variables included in the study. Prior achievement in both reading and mathematics was highly correlated with present achievement. Mathematics achievement had a statistically significant correlation with all predictor variables except for teacher certification type. Reading achievement had a statistically

significant correlation with all predictor variables. The positive correlation between certification type and both mathematics and reading achievement indicated that students studying under teachers with traditional certifications have higher achievement than students studying under teachers with alternative certifications. The negative correlation between degree level and mathematics achievement indicated that students studying under mathematics teachers with bachelor's degrees perform better, on average, than students studying under mathematics teachers with master's degrees, while the positive correlation between degree level and reading achievement indicated that students studying under reading teachers with bachelor's degrees perform worse, on average, than students studying under reading teachers with master's degrees. Certification type was negative correlated with degree level, indicating a trend for more teachers with master's degrees to have alternative certifications. Mathematics and reading achievement were both negatively correlated with FRL and LEP, indicating that students who receive FRL or LEP services perform lower, on average, than their peers who do not receive FRL or LEP services.

Table 4

Correlations and Descriptive Statistics from Mathematics and Reading Achievement Models

	Achievement	PriorAchieve	LEP	FRL	Certification	Degree	Mean	Std. Dev.
Achievement	---	0.765*	-0.243*	-0.273*	0.070*	0.111*	71.690	18.638
Prior Achievement	0.774*	---	-0.267*	-0.278*	0.094*	0.110*	71.010	19.025
LEP	-0.128*	-0.157*	---	0.309*	-0.088*	-0.022	0.200	
FRL	-0.226*	-0.242*	0.309*	---	-0.121*	-0.036*	0.580	
Certification	0.011	0.008	-0.129*	-0.066*	---	-0.099*	0.730	
Degree	-0.036*	-0.031	0.128*	0.097*	-0.071*	---	0.380	
Mean	69.570	70.550	0.200	0.580	0.660	0.280		
Std. Dev.	20.521	19.726						

*Note.** $p < .01$. Correlations below the diagonal are mathematics--related; correlations above the diagonal are reading -related.

Evaluation of VAMs

The mathematics and reading analyses were performed separately, and the results are presented in Tables 5 and 6, respectively. The proportion of variance at each level (i.e., student, teacher, and school) was calculated for mathematics and reading models based on the unconditional VAM as shown in Equations 1a to 1d. For the mathematics and reading models, we found 84% and 76.7% of the total variance to be within classrooms, 16% and 23.1% of the total variance to be among classrooms and within schools, and 0.1% and 0.2% of the total variance to be among schools. These calculations were performed using the formulas provided by Raudenbush and Bryk (2002):

$\frac{\sigma^2}{\sigma^2 + \tau_\pi + \tau_\beta}$	=	the proportion of variance within classrooms
$\frac{\tau_\pi}{\sigma^2 + \tau_\pi + \tau_\beta}$	=	the proportion of variance among classrooms within schools
$\frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta}$	=	the proportion of variance among schools

The detailed results for the five models ran for each academic outcome are summarized in Tables 5 and 6, including fixed and random effects as well as statistical significance for these effects, and model fit information. The percentage of variance explained is calculated by comparing each model with the subsequent model (i.e., fully unconditional model v Model 1, then Model 1 v Model 2, etc.). Here we highlight some important findings. For Model 1, prior achievement was the only covariate used to predict 2018-19 achievement. Approximately 64.5% of the variance in 2018-19 mathematics achievement and 54.8% of the variance in 2018-

19 reading achievement at the student level was explained by adding in prior achievement as a Level 1 predictor. In both models, the fixed effect for prior achievement was positive and significant ($p < .001$).

For Model 2, a student's status as receiving FRL or participating in the district's LEP program were added in as covariates. Approximately 1% additional variance in mathematics achievement and 1% additional variance in reading achievement at the student level was explained by adding in both of these covariates as Level 1 predictors. In both models, the fixed effect for prior achievement and FRL status (i.e., receiving FRL is negatively associated with achievement) was significant ($p < .001$), while the LEP variable was not statistically significant.

In Model 3, a teacher's certification type (traditional or alternative) and highest degree level (bachelors or masters) were added as Level 2 covariates. The percent of variance explained at the teacher's level in the mathematics model was negative, which indicated there was more variance within teachers than between teachers (Snidjers & Bosker, 1999). Approximately 2.8% of variance at the teacher's level was explained in the reading achievement model by adding in these two covariates. In the math model, the fixed effects for both prior achievement and socioeconomic status were again statistically significant ($p < .001$), while in the reading model only the fixed effect for prior achievement was statistically significant. The fixed effects for the LEP variable and the fixed effects for the teacher's degree and certification were not statistically significant for either model.

Model 4 added in aggregated student-level demographics as Level 3 predictors. The addition of these covariates explained 10.9% of variance among teachers in the mathematics model and 3.7% of variance among teachers in the reading model. The addition of these

variables explained 71.7% of variance among schools in the mathematics model and 40.1% of variation among schools in the reading model compared to the previous teacher characteristic model. Prior achievement had a statistically significant ($p < .001$) fixed effect in both mathematics and reading models, while socioeconomic status had a statistically significant fixed effect in the mathematics models only. No other fixed effects had statistical significance.

Finally, Model 5 added in two contrasts coded variables to measure the effect of school characteristics. The addition of these variables explained 3.7% of variance among teachers in the both the mathematics and reading model. Both the mathematics model and the reading model had a negative variance explained for the school level variance, indicating there was more variance within the schools than between the schools (Snidjers & Bosker, 1999). In both the reading and the mathematics model, only the fixed effect for prior achievement had any statistical significance ($p < .001$).

Teacher-level variance was statistically significant across all five models, while school-level variance was not statistically significant across all five models. This is fitting with the intraclass correlations previously calculated, which showed 16% of the total variance in mathematics achievement and 23.1% of the total variance in reading achievement was at the teacher level, while only .1% of the total variance in mathematics achievement and .2% of the total variance in reading achievement to be between schools.

To evaluate model fit, four different model fit indices were calculated: Akaike information criterion (AIC), Bayesian information criterion (BIC), consistent AIC (CAIC), and sample-size adjusted BIC (SABIC). The results are shown at the bottom of Tables 5 and 6 under "model fit."

Table 5

Fixed and Random Effects from Mathematics Achievement Models

Models		1. Base	2. Student Demographics	3. Teacher Characteristics	4. Aggregated Demographics	5. School Characteristics
Student-level fixed effects	Prior mathematics	0.823**	0.815**	0.823*	.0922**	.880**
	FRL		-1.812**	-2.137**	-8.512**	-4.648
	LEP		-0.065	-0.175	8.835	2.702
Teacher-level fixed effects	Certification Type			0.758	11.409171	30.474
	Degree Level			-0.128	10.344261	5.237
School-level fixed effects	Percent FRL				25.448	70.438
	Percent LEP				-18.688	-44.810
	C1					2.371
	C2					20.138
Random Effects	Student-level variance	130.465	129.794	129.719	129.026	128.260
	Teacher-level variance	43.371**	43.446**	43.647**	38.908**	37.454**
	School-level variance	0.029	0.101	0.066	.019	.064
Variance Explained	Student Level	64.5%	.5%	.1%	.5%	.6%
	Teacher Level	38.0%	-.1%	-.5%	10.9%	3.7%
	School Level	91.1%	-251.0%	34.7%	71.7%	-243.3%
Model fit	Deviance	35938.51	35915.43	35913.21	35877.96	35847.44
	Estimated parameters	5	7	15	39	63

(table continues)

Models		1. Base	2. Student Demographics	3. Teacher Characteristics	4. Aggregated Demographics	5. School Characteristics
	AIC	35948.51	35929.43	35943.21	35955.96	35973.44
	BIC	35980.71	35974.51	36039.80	36207.10	36379.13
	CAIC	35985.71	35981.51	36054.80	36246.10	366442.13
	SABIC	35964.82	35952.27	35992.14	36083.17	36178.94

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

Table 6

Fixed and Random Effects from Reading Achievement Models

Models		1. Base	2. Student Demographics	3. Teacher Characteristics	4. Aggregated Demographics	5. School Characteristics
Student-level fixed effects	Prior mathematics	.732**	.719**	.728**	.872**	.739**
	FRL		-1.890**	-.238	-2.906	5.025
	LEP		-.772	-1.136	4.757	10.286
Teacher-level fixed effects	Certification Type			1.470	7.504	4.529
	Degree Level			2.258	17.849	.772
School-level fixed effects	Percent FRL				16.278	15.579
	Percent LEP				12.251	12.047
	C1					1.967
	C2					-5.255
Random Effects	Student-level variance	123.336	122.610	122.451	121.566	120.905

(table continues)

Models		1. Base	2. Student Demographics	3. Teacher Characteristics	4. Aggregated Demographics	5. School Characteristics
	Teacher-level variance	21.423**	20.421**	19.847**	19.113**	18.403**
	School-level variance	.033	.028	.027	.016	.0242
Variance Explained	Student Level	54.8%	.6%	.1%	.7%	.5%
	Teacher Level	73.9%	4.7%	2.8%	3.7%	3.7%
	School Level	95.5%	14.1%	4.9%	40.1%	-50.9%
Model fit	Deviance	34775.88	34745.06	34736.52	34700.77	34673.02
	Estimated parameters	5	7	15	39	63
	AIC	34785.88	34759.06	34766.52	34778.77	34799.03
	BIC	34817.96	34803.97	34862.74	35028.95	35203.17
	CAIC	34822.96	34810.97	35028.95	35067.95	35266.17
	SABIC	34802.07	34781.72	34815.08	34905.03	35002.98

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

Because these represent the degree of inaccuracy, or "badness of fit," in the model, the model with the lowest number for these indices is considered to be the best model (McCoach & Black, 2008). All four model fit indices indicated the student demographics model as the best fitting model. This is consistent with the findings of the fixed effects, as the student-level covariates were the only statistically significant fixed effects. McCoach and Black (2008) cautioned while model fit indices are a "prudent course of action (p. 261)" for evaluating model fit, the human judgement of the researcher should also be taken into consideration. Henson (2006) also noted, not only is statistical significance important, but practical and clinical significance may have important implications as well. Because these VAMs are often used to make high-stakes decisions that could impact a teacher's livelihood, it is worth examining all models further.

Correlations between VAMs

To investigate variations in the estimates of teacher and school value-added scores using the five different models, Spearman rank correlations were calculated among both the teacher-level residuals and the school-level residuals between the simpler model and the ensuing more complex models (see Table 7). For both teacher- and school-level residuals, the correlations between the simpler model and the ensuing models' VAM scores become weaker as more covariates were included, indicating less similarities in value-added scores as the models become more complex. Both reading and math teacher-level and school-level VAM scores show strong correlations between models; however, the correlations overall are stronger for the teacher value-added scores than for the school value-added scores. These high correlations could be contributed to the use of EB estimates rather than OLS estimates, which shrink or pull the residuals to the mean (Hox, 2002). Because the teacher clusters had a smaller

number of student scores than the school clusters, their residuals will be more greatly impacted by the shrinkage to the mean that occurs with EB estimation.

Teacher-level value-added rankings in mathematics are more highly correlated than those in reading between the first three models, while the teacher-level value-added rankings in reading tend to be more highly correlated than those in math between the last two models (see Table 7). At the school-level, however, the correlations between the school-level rankings in the math models are almost always higher than those in the reading models.

Pearson correlations between the teacher- and school-level residuals (i.e., value-added scores) and the included student demographics were calculated to evaluate whether some of the models may be more sensitive to the student demographics than others (see Table 8). Not surprisingly, for school-level residuals, the models that control for student demographics (i.e., aggregated demographics and school characteristics) do not have statistically significant correlations with either of the student demographics, while models that do not control for student demographics do have statistically significant correlations with those demographics. For the teacher-level residuals, however, the models that controls for student demographics often had statistically significant correlations with student demographics. Across all models, the correlations between student demographics and the teacher-level residuals are stronger for the reading model than for the mathematics model, indicating teachers of reading are more likely to be ranked lower if they have higher percentages of students who come from low socioeconomic households or are enrolled in the district's LEP program.

Table 7

Spearman Rank Correlations between Value-Added Rankings across Models

Models	1. Base		2. Student Demographics		3. Teacher Characteristics		4. Aggregated Demographics		5. School Characteristics	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
Teacher VAM										
1. Base	1.00	1.00	.998*	.995*	.998*	.986*	.966*	.960*	.938*	.954*
2. StuDem			1.00	1.00	.999*	.989*	.962*	.964*	.933*	.957*
3. TchCh					1.00	1.00	.961*	.979*	.932*	.972*
4. AggDem							1.00	1.00	.966*	.991*
5. SchlCh									1.00	1.00
School VAM										
1. Base	1.00	1.00	.996*	.939*	.987*	.908*	.961*	.831*	.920*	.744*
2. StuDem			1.00	1.00	.994*	.963*	.946*	.873*	.904*	.794*
3. TchCh					1.00	1.00	.928*	.916*	.889*	.827*
4. AggDem							1.00	1.00	.889*	.955*
5. SchlCh									1.00	1.00

*Note.** $p < .01$

Table 8

Correlations between Value-Added Rankings and Student Demographics

	FRL		LEP	
	Math	Reading	Math	Reading
Teacher VAM				
1. Base	.004	-.228*	.022	-.202*
2. Student Demographics	.058*	-.139*	.062*	-.125*
3. Teacher Characteristics	.066*	-.140*	.076*	-.135*
4. Agg Demographics	-.029**	-.094*	-.009	-.142*
5. School Characteristics	-.029	-.090*	-.011	-.145*
School VAM				
1. Base	.179*	-.497*	.183*	-.244*
2. Student Demographics	.316*	-.188*	.279*	.051*
3. Teacher Characteristics	.331*	-.180*	.310*	.008
4. Agg Demographics	.001	-.028	.018	-.017
5. School Characteristics	.019	-.022	.017	-.022

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

Evaluation of Teacher and School Rankings

To investigate the changes in teacher and school ranking across the five models, teacher and school residuals for each model were sorted in ascending order of number and divided into quintiles. Several authors have used various breaking points to group teachers, including quartiles (Hawley et al., 2017), quintiles (Leckie, 2018; Palardy & Peng, 2015; Parsons et al., 2019; Schochet & Chiang, 2013), or deciles (Newton et al., 2010). Because the majority of the literature used quintiles as the grouping system for ranking teacher value-added scores, quintiles were chosen for the current study. A teacher or school's quintile placement can be interpreted as their performance level. For example, a teacher or school in the 5th quintile has a

value-added score higher than 80% of the teachers or schools in the lower four quintiles. The following section looks at the changes across quintiles by first examining the number and percent of changes in quintile ranking between a given model and the subsequent model, then examining the total changes in quintile ranking for a given teacher or school, and finally, examining the number and percent of changes in quintile ranking from the base model to the other four models.

First, after the quintiles were created for each model, the number of teachers or schools whose quintile ranking changed between models was counted (see Table 9). If a teacher was in the 2nd quintile in Model 1, but ranked in the 3rd quintile in Model 2, that change was counted towards the number of changes from Model 1 (i.e., the previous model) to Model 2 (i.e., the consecutive model). For example, the top left entry of 6(5.8%) indicates that six teachers (5.8% of all teachers) changed ranking by one quintile from Model 1 to Model 2. The largest number of teacher-level quintile rank changes for both subjects occurred from Model 3 to Model 4, with 14 math teachers (12.2% of all math teachers) and 20 reading teachers (17.2% of all reading teachers) changing one quintile ranking between these two models. At the school level, the number of quintile rank changes from model to model remained fairly consistent across the mathematics models. The school-level quintile rank changes for reading, however, changed quite a bit more from Model 1 to Model 2, with 15% and 45% of schools changing in quintile rank in mathematics and reading, respectively, and again from Model 4 to Model 5, with 25% and 40% of schools changing in quintile rank in mathematics and reading, respectively.

Table 9

Number and Percent of Changes in Quintile Rankings

	2. Student Demographics		3. Teacher Characteristics		4. Aggregated Demographics		5. School Characteristics	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading
Teacher VAM								
One	6(5.8%)	10(8.6%)	4(3.9%)	14(12.1%)	14(13.6%)	20(17.2%)	16(15.5%)	14(12.1%)
Two	(0%)	(0%)	(0%)	(0%)	1(1.0%)	(0%)	2(1.9%)	(0%)
School VAM								
One	3(15%)	9(45%)	5(25%)	5(25%)	5(25%)	4(20%)	5(25%)	8(40%)
Two	(0%)	(0%)	(0%)	(0%)	2(10%)	1(5%)	2(10%)	(0%)

Note. Entries are in the *n* (%) format. "One" represents the number and percentage of teachers or schools who moved one quintile from the previous model, "two" represents the number and percentage of teachers or schools who moved two quintiles from the previous model.

Frequently, a teacher or school would change one or two quintile rankings from the previous model, but in a subsequent model would change in the opposite direction, or even back to the original ranking. For example, one math teacher's residual/valued added score began in the fourth quintile in Model 1, remained in the fourth quintile in Model 2, changed to the fifth quintile in Model 3, then moved to the third quintile in Model 4 and stayed there in Model 5. A single teacher then could account for two or more quintile changes in Table 9. Another table is needed to better understand the changes in quintile ranking of a single teacher or school across all models.

To better display the changes made by a single teacher or school across all models as well as the number of teachers or schools who showed no changes in quintile ranking across all models, Table 10 was created. To continue with the example of the math teacher from above, this teacher was in the fifth quintile in their highest ranking and the third quintile in their lowest ranking. Because this is a difference of two quintile rankings, this teacher would appear in the row "Two" under the math column.

In the teacher-level results, 32% of mathematics teachers and 37% of reading teachers changed in ranking by at least one quintile across the five models. These results are consistent with the results from Schochet and Chiang's (2013) study, which found teacher misclassifications as being either high-performing or low-performing when in reality they were not, was likely to occur 26% of the time. The school-level quintile rankings were more drastic, with 55% and 65% of schools changing in ranking by at least one quintile in mathematics and reading, respectively. One school's quintile rank in reading changed over the gamut of possibilities, with a difference of 4 quintile rankings across the models.

Table 10

Number and Percent of Changes for a Single Teacher or School across Models

	Math	Reading
Teacher VAM		
Zero	70 (68%)	73(62.9%)
One	28(27.2%)	43(25.9%)
Two	5(4.9%)	0(0%)
School VAM		
Zero	9(45%)	7(35%)
One	7(35%)	10(50%)
Two	4(20%)	2(10%)
Three	(0%)	(0%)
Four	(0%)	1(5%)

Note. Entries are in the *n* (%) format. "One" represents the number and percentage of teachers or schools that moved one total quintile across all model specifications, "two" represents the number of teachers or schools that moved two total quintiles across all model specifications.

Finally, it is helpful to view the changes of each individual model from the original base model. Similar to Table 10, Figure 1 displays the changes in a single teacher or a single school's quintile ranking. Unlike Table 10, the changes in quintile rankings for a given model displayed in Figure 1 are always calculated from the original quintile ranking in the base model. Additionally, this figure takes into consideration the direction of the change in quintile ranking. The math teacher previously described is used again as an example. This teacher was in the fourth quintile in Model 1. For the four subsequent models, they were in the fourth quintile, the fifth quintile, the third quintile, and the third quintile. This is shown in Figure 1, the math teacher

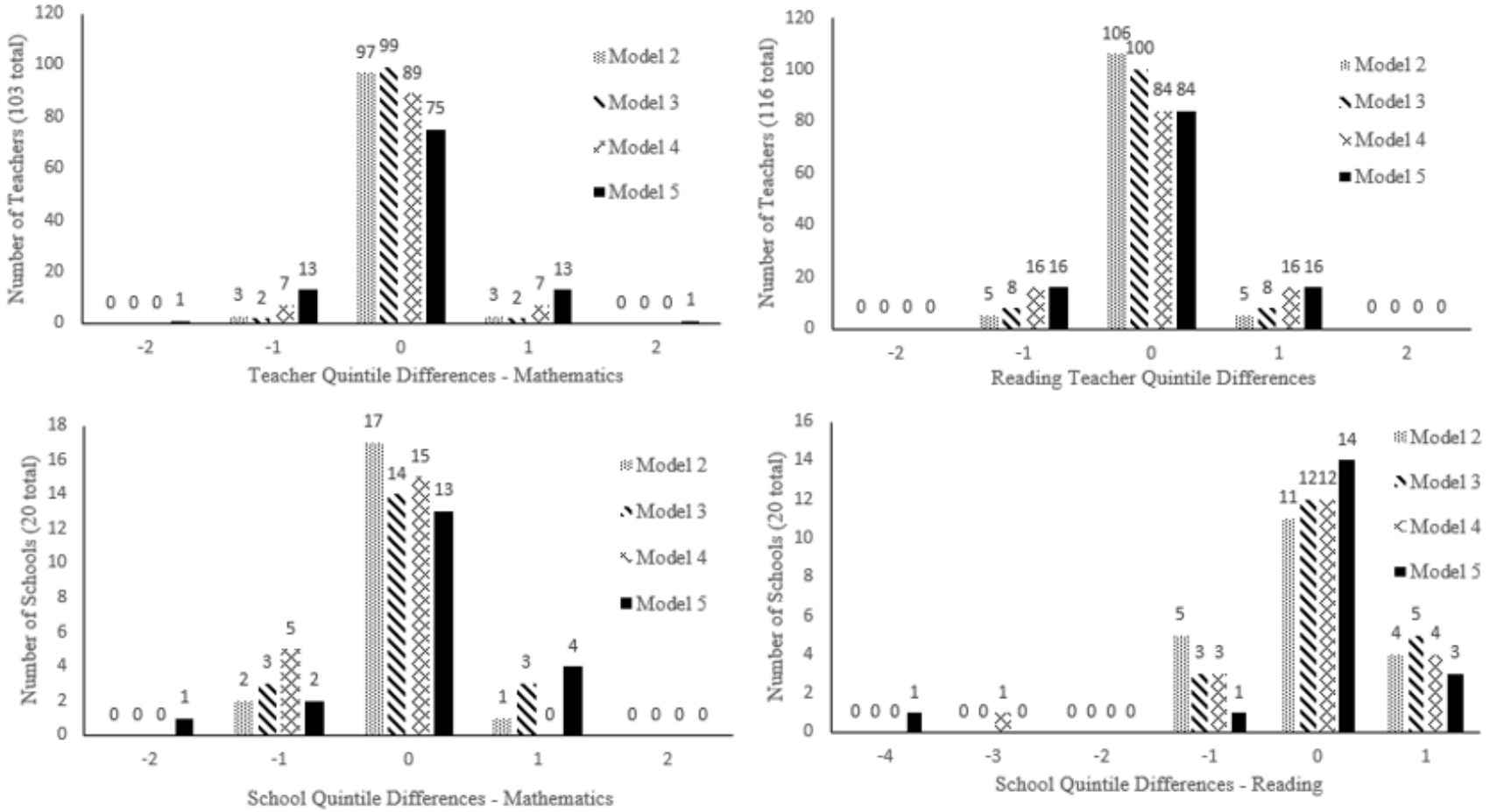
quintile differences, as one of the teachers who had a 0 difference for Model 2, a +1 difference for Model 3, and a -1 difference for both Models 4 and 5.

Figure 1 shows the majority of teachers showed no changes in quintile rankings from Model 1 to each of the other four models. For math teachers, the largest change in ranking from the Model 1 to Model 5, with 13 teachers moving up one quintile and 13 teachers moving down one quintile. Reading teachers had 16 teachers moving up one quintile from Model 1 and 16 teachers moving down one quintile from Model 1 in both Model 4 and Model 5. The symmetry of both the math and reading teacher quintile differences shows that the number of teachers who changed from a lower-performing quintile to a higher-performing quintile in each model is approximately the same as the number of teachers changed from a higher-performing quintile to a lower performing quintile in each model.

The changes in the school-level quintile rankings from the base model to each of the other four models show more disparity than the teacher-level changes in quintile ranking. The changes in ranking are not symmetric for either the mathematics or the reading models. In the mathematics models, more schools show a change in the negative direction than the positive direction. In the reading models, the number of schools that change ranking in the positive direction is approximately the same as the number of schools that change ranking in the negative direction, however, there is a greater disparity in the change in ranking in the negative direction, with one school moving down three quintiles and another school moving down four quintiles, or the entire range of possible rankings.

Figure 1

Quintile Rank Differences from the Base Model to Each Alternative Model



Note. Models 2 – 5 are, in order: "Student Demographics," "Teacher Characteristics," "Aggregated Demographics," and "School Characteristics." "-2" indicates a decrease of two quintiles, "-1" indicates a decrease of one quintile, "0" indicates no change, "1" indicates an increase of one quintile, and "2" indicates an increase of two quintiles.

Discussion

Across all models, the only statistically significant fixed effects are related to students' prior math achievement and their socioeconomic status, however, many changes occur in both teacher and school value-added score quintile rankings as other variables are controlled for. Additionally, the correlations between the value-added score rankings for both teachers and schools become weaker as additional covariates are included in the model, indicating more and more of a lack of consistency in rankings at both the teacher- and school- level. So, while the inclusion of some covariates may not produce statistically significant fixed effects, there are noticeable changes in the rankings for individual teachers and/or schools as a result of controlling for these covariates. Teachers affected by any high-stakes decisions, such as contract renewals, made from VAM applications may find these changes personally significant. In the following sections, I discuss the implications of these results in terms of the research questions.

Impact of Including Teacher-Level Covariates on Teacher Value-Added Scores

The results of this study indicate that the teacher level covariates included have a larger impact on reading teacher value-added scores than mathematics teachers value-added scores. When certification type and degree level are added into the model, 4% of the mathematics teachers and 12% of the reading teachers change one quintile (see Table 9). This is consistent with the correlations found in Table 4. The certification type and degree level have statistically significant correlations with reading achievement; however, only degree level has a statistically significant correlation with mathematics achievement. For reading achievement, the correlations indicate a statistically significant, positive relationship with both certification type

and degree level, while for mathematics achievement the correlations indicate a non-significant, positive relationship with certification type and a statistically significant, negative relationship with degree level. The mathematics results are inconsistent with past research, which has found students may have higher achievement when studying under teachers with traditional certifications or a higher degree level (Darling-Hammond et al, 2005; Irvine, 2019). A possible explanation for this could be that the teacher with the higher degree level has more content knowledge in mathematics which translates into that teacher speaking at too high of a cognition level during class or having trouble understanding the misconceptions of struggling learners. It is also important when analyzing the negative direction of this correlations to consider the effect size. While the correlation was statistically significant, these results came from a very large sample size. The r-squared type effect size for the correlation between degree level and math achievement is only .0013, indicating only .13% of the variance in math achievement is accounted for by a teacher's degree level.

The certification type and degree level have more of an impact on the reading teacher's value-added scores than the mathematics teacher's value-added scores, but overall, the inclusion of these two covariates in the third model have a small impact compared to the other models. This model has the smallest number of changes in mathematics teacher's value-added score quintile ranking, and the second smallest number of changes in reading teacher's value-added score quintile ranking. The two covariates included, certification type and degree level, do not speak to that teacher's motivation, self-efficacy, or other personality traits, which were not included in this study. The impact of certification type and degree level on teacher value-added scores from this study are enough to merit future research that includes latent traits

such as teacher motivation or self-efficacy as teacher level covariates.

Impact of Including School-Level Covariates on Teacher Value-Added Scores

Aggregated School-Level Covariates

The results of this study indicate that inclusion of student-level demographics aggregated at the school level has a large impact on both reading and mathematics teacher's value-added scores. When FRL and LEP are added as aggregated demographics at the school level, 14% of mathematics teachers and 17% of reading teachers change ranking by one quintile from the previous model, and 1% of mathematics teachers change ranking by two quintiles from the previous model. This is the highest percentage of change from the previous model in reading teachers and the second highest percentage of change from the previous model in mathematics teachers (see Table 9).

More prevalent and stronger negative correlations exist between student demographics and reading teacher's value-added scores than mathematics teacher's value-added scores across all models (see Table 8). These stronger negative correlations are indicative of the stronger impact of these variables on the reading teachers' value-added scores. As a reading teacher working in a school with a higher percentage of students who received FRL or a higher percentage of students who are enrolled in the district's LEP program, that teacher is more likely to be ranked lower compared to their peers in terms of their value-added scores. The results in Tables 9 and 10 agree. Table 9 confirms as student demographics are controlled for first in Model 2 and again in Model 4, there are more reading teachers changing quintiles than mathematics teachers, with the addition of the aggregated demographics covariates in the

fourth model resulting in the highest percentage of reading teachers changing in quintile ranking.

The larger impact on reading teacher value-added scores compared to mathematics teacher value-added scores is partially explained by the use of a child's English language proficiency as a predictor variable. Because of the universal nature of numbers, we would expect students who are learning English as a second language to struggle more in reading than in mathematics. Therefore, it is very plausible that reading teachers will be more impacted by their student's English language proficiency than their mathematics counterparts. This may also be explained by the use of socioeconomic status as a predictor variable, although the explanation is not as direct. Educational literature is very consistent regarding the impact of socioeconomic status and parental involvement on a child's early literacy (Froiland et al., 2013; Hemmerechts et al., 2017; Kuhl, 2011; Waldfogel, 2012). Parents with higher socioeconomic resources are more likely to engage in behaviors such as reading with their children and speaking to their children using elevated vocabulary (Hemmerechts et al., 2017; Hart & Risley, 1995; Rodriguez et al., 2009; Rodriguez & Tamis-LeMonda, 2011). The literature on the impact of socioeconomic status and parental involvement on early numeracy is not as abundant, but does point towards a similar pattern (Park & Holland, 2017). A possible explanation for the discrepancy found here may be that, while these parents with a higher socioeconomic status are more likely to read to their children every night, they may not be going out of their way to discuss mathematics as often. It is plausible that while children may revel in bedtime stories, bedtime mathematics may not be quite as popular!

School Characteristics – SOC Variables

The results of this study indicate that the inclusion of the SOC variables at the school level had a large impact on both reading and mathematics teacher's value-added scores. when these contrast coded variables were included in the school-level model, 16% of mathematics teachers and 12% of reading teachers changed one quintile ranking from the previous model, and 2% of mathematics teachers changed two quintiles rankings from the previous model. This was the highest percentage of mathematics teachers changing in quintile ranking and the second highest percentage of reading teachers changing in ranking. This indicates mathematics teachers may be more heavily impacted by the SOC programs offered in this school district. Nine of the twelve schools offering SOC programs have the Suzuki Strings programs, and prior research has found there is often a positive correlation between music programs or music ability and mathematics achievement (Cranmore & Tunks, 2015; Hattie, 2018; Southgate & Roscigno, 2009).

The large number of changes in both reading and mathematics teacher value-added scores can be explained by the additional parental involvement required for school of choice programs. Parents who enrolled their children in the Spanish Immersion program must sign a contract stating they will read with their children for a minimum of 20 minutes a night in English, and parent who enrolled their children in the Suzuki Strings program must attend one practice a week with their child. Prior research has found that parental involvement not only creates a culture of academic responsibility, but can moderate the effects of socioeconomic status (Mwangi et al., 2019).

Impact of Including School Level Covariates on School Value-Added Scores

Aggregated School-Level Covariates

The results of this study indicate that inclusion of student-level demographics aggregated at the school level has a large impact on a school's value-added score in both reading and mathematics. When FRL and LEP are added as aggregated demographics at the school level, 25% and 20% of schools' value-added score in mathematics and reading, respectively, change in ranking by one quintile from the previous model, and 10% and 5% of schools' value-added score in mathematics and reading, respectively, change in ranking by two quintiles from the previous model. This change in number and percent of quintile rank changes is tied with Model 5 as the greatest number of changes in school quintile ranking in mathematics across all models (see Table 9).

There is a positive correlation between school-level residuals and the percentage of students receiving both FRL and LEP services in the mathematics models (see Table 8). A student's FRL and LEP status is controlled for in Model 2 and the school's percent of students receiving FRL and LEP services is controlled for again in Model 4. This indicates that as the influence of these variables are removed, these schools with higher percentages of kids receiving FRL and LEP services will have a higher value-added score in mathematics. Relating these residual scores back to student achievement, this indicates schools with higher percentages of kids receiving FRL or LEP services are showing more improvement, regardless of prior achievement, than the schools with lower percentages of kids receiving FRL or LEP services, and therefore would have a higher value-added score.

School Characteristics – SOC Variables

The results of this study indicate that inclusion of the SOC variables has a large impact on a school's value-added score in both reading and mathematics. When the contrast codes for these variables are included at the school level, 25% and 40% of schools' value-added score in mathematics and reading, respectively, change in ranking by one quintile from the previous model, and 10% schools' value-added score in mathematics change in ranking by two quintiles from the previous model. This change in number and percent of quintile rank changes is tied with Model 4 as the greatest number of changes in school quintile ranking in mathematics and is the second greatest number of changes in school quintile ranking in reading across all models (see Table 9). Prior research indicates that school programs may have an impact on student achievement (Hattie, 2018). This study adds to the literature by finding that this will in turn also impact the school's value-added score.

Implications

Implications for school administrators and policymakers based on the results of this study indicate that VAMs should be used with caution when making high stakes decisions based on results from VAM data. If VAMs are going to be used to evaluate teacher or school value-added scores, student level covariates should be controlled for at the student level, and aggregated student level covariates should be controlled for at the school level. Because of the inconsistency in teacher and school value-added scores, administrators and policy makers should also take into account multiple years of value-added scores to ensure the teacher or school is consistently shown to be low- or high- performing.

Limitations and Future Research

Limitations of this study include the sample and the covariates chosen. All data was collected from one school district and only included Grades 4 - 6, so results may not generalize to other grade levels or to school districts without similar characteristics as this school district. In addition, the teacher level covariates and school level SOC covariates chosen did not have statistically significant fixed effects across the models in which they were included. The teacher degree level and certification type do not explain other traits a teacher may have, such as motivation and self-efficacy. Future studies should attempt to measure these as latent variables and include those latent variables as teacher level covariates. Hattie's (2018) meta-analysis found school of choice programs such as these to be likely to have a small positive impact on student achievement, but found several other school characteristics, such as school size, after-school programs, and service learning programs, to be more likely to have a stronger impact on student achievement. Future studies should attempt to include in their models other variables such as these described by Hattie.

Summary and Conclusions

The purpose of the current study was to investigate the impact of including teacher-and school-level covariates on teacher and school value-added scores. This study found the addition of covariates at both the teacher and the school level has an impact on the quintile ranking of both the teacher and the school's value-added score, with school-level covariates having the largest impact on both teacher and school value-added scores. These results are consistent with the findings of Schochet and Chiang (2013), who found that VAMs that do not include covariates are likely to misclassify approximately 26% of teachers included in the model as

either high or lower performing. In this study, 32% of mathematics teachers and 37% of reading teachers changed quintile ranking at least once depending on the model used, and 55% of school rankings in mathematics and 65% of school rankings in reading changed quintile ranking at least once depending on the model used. With VAMs often being used for high-stakes decisions such as teacher pay, teacher contract renewal, or school accountability (Braun, 2005; Chetty et al., 2014; Kane et al., 2008; Kane et al., 2013), this is a large number of individuals who may have life-altering experiences based on decisions for inaccurate VAM results.

References

Amrein-Beardsley, A. (2019). The education value-added assessment system (EVAAS) on trial: A precedent-setting lawsuit with implications for policy and practice. *JEP: eJournal of Education Policy, 37*(2), 65–75. https://doi.org/10.3102/0013189X08316420

Atteberry, A., & Mangan, D. (2020). The sensitivity of teacher value-added score to the use of fall or spring test scores. *Educational Researcher, 49*(5), 335–349. https://doi.org/10.3102/0013189X20922993

Baker, M., & Johnston, P. (2010). The impact of socioeconomic status on high stakes testing reexamined. *Journal of Instructional Psychology*, *37*(3), 193–200. https://www.academia.edu/23719708/The_Impact_of_Socioeconomic_Status_on_High_Stakes_Testing_Reexamined

Basileo, L. D., & Toth, M. (2019). A state level analysis of the Marzano Teacher Evaluation Model: Predicting teacher value-added measures with observation scores. *Practical Assessment, Research & Evaluation, 24*(6), 1–14. https://doi.org/10.7275/cc5b-6j43

Bianchi, A. J., & Lancianese, D. A. (2005). No Child Left Behind?: Role/identity development of the "good student". *International Journal of Educational Policy, Research, and Practice: Reconceptualizing Childhood Studies*, *6*(1), 3–29. https://files.eric.ed.gov/fulltext/EJ795130.pdf

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* Guilford Press.

Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. *Educational Testing Service*.

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

Chetty, R., Friedman, J., & Rockoff, J. (2014). Discussion of the American Statistical Association's statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy, 1*(1), 111–113. https://doi.org/10.1080/2330443X.2014.955227

Chmielewski, A. K., & Reardon, S. F. (2016). Patterns of cross-national variation in the association between income and academic achievement. *AERA Open*, *2*(3), 1–27. https://doi.org/10.1177/2332858416649593

Cranmore, J., & Tunks, J. (2015). Brain research on the study of music and mathematics: A meta-synthesis. *Journal of Mathematics Education*, *8*(2), 139-157.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives*, *13*(42). https://doi.org/10.14507/epaa.v13n42.2005

Engin, G. (2020). An examination of primary school students' academic achievements and motivation in terms of parents' attitudes, teacher motivation, teacher self-efficacy and leadership approach. *International Journal of Progressive Education*, *16*(1), 257–276. https://doi.org/10.29329/ijpe.2020.228.18

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

Froiland, J. M., Powell, D. R., Diamond, K. E., & Son, S.-H.C. (2013). Neighborhood socioeconomic well-being, home literacy, and early literacy skills of at-risk preschoolers. *Psychology in the Schools, 50(8)*, 755 – 769.

Goldhaber, D., & Theobald, R. (2012). Do different value-added models tell us the same things? What we know series: Value-added methods and applications. Knowledge Brief 4. *Carnegie Foundation for the Advancement of Teaching*.

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. https://doi.org/10.1080/2330443X.2013.856169

Gonzalez, K., & Maxwell, G. M. (2018). Mathematics teachers' efficacy, experience, certification and their impact on student achievement. *Journal of Instructional Pedagogies*, *21*.

Hamilton, R., McCoach, D. B., Tutwiler, M. S., Siegle, D., Gubbins, E. J., Callahan, C. M., Brodersen, A. V., & Mun, R. U. (2018). Disentangling the roles of institutional and individual poverty in the identification of gifted students. *Gifted Child Quarterly*, *62*(1), 6-24. https://doi.org/10.1177/0016986217738053

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management, 28*(4), 693–699. https://doi.org/10.1002/pam.20464

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hattie, J. (2018, March 28). *Hattie ranking: 250 influences and effect sizes related to student achievement*. Corwin Visible Learning Plus. https://www.visiblelearningmetax.com/

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

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration, 47*(2), 227–249. https://doi.org/10.1108/09578230910941066

Hemmerechts, K., Agirdag, O., & Kavadias, D. (2017). The relationship between parental literacy involvement, socio-economic status and reading literacy. *Educational Review*, *69*(1), 85-101.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, *34*(5), 601–629. https://doi.org/10.1177/0011000005283558

Heyneman, S. P. (2005). Student background and student achievement: What is the right question? *American Journal of Education*, *112*(1), 1–9. https://doi.org/10.1086/444512

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*(4), 700–712. https://doi.org/10.1002/pam.20463

Houston Independent School District. (2018). Pre-exit English learner student performance English STAAR and TELPAS 2017-18. *HISD Research and Accountability Analyzing Data, Measuring Performance.*

Houston Independent School District. (2019). English as a second language student performance report English STAAR and TELPAS 2018-19. *HISD Research and Accountability Analyzing Data, Measuring Performance.*

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Lawrence Erlbaum Associates.

IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp

Irvine, J. (2019). Relationship between teaching experience and teacher effectiveness: Implications for policy decisions. *Journal of Instructional Pedagogies*, *22*.

Kane, T. J., McCaffrey, T. M., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Bill and Melinda Gates Foundation.

Kane, T. J., Staiger, D. O., & National Bureau of Economic Research (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER working paper no. 14607. In *National Bureau of Economic Research.* National Bureau of Economic Research.

Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. Routledge.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18–42. https://doi.org/10.1162/EDFP_a_00027

Kuhl, P. K. (2011). Early language learning and literacy: Neuroscience implications for education. *Mind, Brain, and Education, 5(3), 128 – 142.*

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67. https://doi.org/10.1111/j.1745-3984.2007.00026.x

Leckie, G. (2018). Avoiding bias when estimating the consistency and stability of value-added school effects. *Journal of Educational and Behavioral Statistics, 43*(4), 440–468. https://doi.org/10.3102/1076998618755351

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology, 81*(1), 59–77. https://doi.org/10.1348/000709910X503501

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational Behavioral Statistics, 29*(1), 67–101. https://doi.org/10.3102/10769986029001067

McCoach, D. B., & Black, A. C. (2008). *Multilevel Modeling of Educational Data*. Information Age Publishing.

McGown, J. A., & Slate, J. R. (2019). Differences by economic status in Grade 3 reading performance: A Texas multiyear study. *Athens Journal of Education*, *6*(3), 189–207. https://doi.org/10.30958/aje.6-3-2

Miley, S. K., & Farmer, A. (2017). English language proficiency and content assessment performance: A comparison of English learners and native English speakers achievement. *English Language Teaching*, *10*(9), 198–207. https://doi.org/10.5539/elt.v10n9p198

Morganstein, D., & Wasserstein, R. (2014). ASA statement of value-added models. *Statistics and Public Policy, 1*(1), 108–110. https://doi.org/10.1080/2330443X.2014.956906

Mwangi, C. A. G., Cabrera, A. F., & Kurban, E. R. (2019). Connecting school and home: Examining parental and school involvement in readiness for college through multilevel SEM. *Research in Higher Education*, *60*(4), 553-575. https://doi.org/10.1007/s11162-018-9520-4

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives,18*(23). https://doi.org/10.14507/epaa.v18n23.2010

Office of Head Start. (2020). *Head start programs.* U.S. Department of Health & Human Services. https://www.acf.hhs.gov/ohs/about/head-start

Palardy, G. & Peng, L. (2015). The effect of summer on value-added assessments of teacher and school performance. *Educational Policy Analysis Archives, 23*(92). http://dx.doi.org/10.14507/epaa.v23.1997

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163–193. https://doi.org/10.3102/0002831210362589

Parsons, E., Koedel, C., & Tan, L. (2019). Accounting for student disadvantage in value-added models. *Journal of Educational and Behavioral Statistics, 44*(2), 144–179. https://doi.org/10.3102/1076998618803889

Ratcliff, N. J., Pritchard, N. A., Knight, C. W., Costner, R. H., Jones, C. R., & Hunt, G. H. (2014). The interaction of school organization and classroom dynamics: Factors impacting student achievement. *Journal of Research in Education*, *24*(2), 3–17. https://files.eric.ed.gov/fulltext/EJ1098174.pdf

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121–129. https://doi.org/10.3102/10769986029001121

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). SAGE Publications.

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

Rodriguez, E. T., & Tamis-LeMonda, C. S. (2011). Trajectories of the home learning environment across the first 5 years: Associations with children's vocabulary and literacy skills at prekindergarten. *Child Development*, *82*(4), 1058-1075.

Rodriguez, E. T., Tamis-LeMonda, C. S., Spellmann, M. E., Pan, B. A., Raikes, H., Lugo-Gil, J., & Luze, G. (2009). The formative role of home literacy experiences across the first three years of life in children from low-income families. *Journal of Applied Developmental Psychology*, *30*(6), 677-694.

Rose, R., Henry, G. T., & Lauen, D. L. (2012). Comparing value added models for estimating teacher effectiveness: Technical briefing. *Carolina Institute for Public Policy. Consortium for Educational Research and Evaluation-North Carolina.* http://publicpolicy.unc.edu/files/2014/02/EE_VAM_Briefing_2-7-12.pdf

Sanders, W. L. & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311. https://doi.org/10.1007/BF00973726

Schochet, P. Z. & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*(2), 142–171. https://doi.org/10.3102%2F1076998611432174

Snidjers, T. & Bosker, R. (1999). Modeled variance in two-level models. *Sociological Methods and Research*, *22*(3), 342 – 363.

Southgate, D. E., & Roscigno, V. J. (2009). The impact of music on childhood and adolescent achievement. *Social Science Quarterly, 90(1),* 4-21. https://doi.org/10.1111/j.1540-6237.2009.00598.x

Statistical Analysis System Education Value-Added Assessment System. (2016). *SAS EVAAS for K-12 statistical models.* https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/sas-evaas-k12-statistical-models-107411.pdf

Texas Education Agency. (2013). *Vertical scale technical report*. https://tea.texas.gov/sites/default/files/2013-STAARVerticalScaleTechReport.pdf

Texas Education Agency. (2019). *2019 accountability manual*. https://tea.texas.gov/texas-schools/accountability/academic-accountability/performance-reporting/2019-accountability-manual

Texas Education Agency. (2020, October 15). *PEIMS data standards*. Retrieved October 15, 2020 from http://ritter.tea.state.tx.us/peims/standards/weds/index.html?c054

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics 29*(1), 11–36. https://doi.org/10.3102%2F10769986029001011

TVAAS. (2019). Statistical models and business rules of TVAAS analysis. *TVAAS Tennessee*. https://tvaas.sas.com/support/TVAAS-Statistical-Models-and-Business-Rules.pdf

Waldfogel, J. (2012). The role of out-of-school factors in the literacy problem. *The Future of Children*, *2*(22), 39–54. https://doi.org/ 10.1353/foc.2012.0016

Whiteside, K. E., Gooch, D., & Norbury, C. F. (2017). English language proficiency and early school attainment among children learning English as an additional language. *Child Development*, *88*(3), 812–827. https://doi.org/10.1111/cdev.12615

Zoda, P., Combs, J. P., & Slate, J. R. (2011). Elementary school size and student performance: A conceptual analysis. *International Journal of Educational Leadership Preparation*, *6*(4), 1–20. https://files.eric.ed.gov/fulltext/EJ974350.pdf

APPENDIX

EXTENDED LITERATURE REVIEW

Evaluation of Model Specification

Many researchers have found the inclusion or lack of inclusion of covariates at the student, teacher, and school level to have a notable effect on the value-added outcomes (Heck, 2009; Palardy & Peng, 2015; Tekwe et al., 2004). These covariates become even more important in models that do not include prior achievement, or lagged achievement (Koedel et al., 2015). In addition to covariates, model choice, estimation type, and effect type, random or fixed, may impact the outcomes at both the teacher and school level (Leckie, 2018; McCaffrey et al., 2014, Tekwe et a.l, 2004; Goldstein et al., 2007; Grady & Beretvas, 2010). Several specific findings, as well as suggestions for future research, are detailed in the sections to follow.

Nuttall, Goldstein, Prosser, and Rasbash (1989) used a multilevel model to study the effectiveness of different schools. They included several factors in their model, including sex, ethnicity, designation as a mixed or single sex school, and designation as a public or private school, in addition to student prior achievement. They concluded an overall concept of school effectiveness was not useful, but rather it would be more meaningful to analyze differences in effectiveness by different subpopulations of the schools. This is a very early study compared to most VAM literature and includes two school-level covariates. Since this study, there have been few studies in the literature regarding school level covariates, although it has been suggested that this should be studied further (Basileo & Toth, 2019; Heck, 2009; Schochet & Chiang, 2013).

Palardy and Peng (2015) used a 3-level hierarchical linear model to determine the effects of summer on value-added models. By comparing year to year testing with fall to spring testing, they found the summer period, when students are not receiving formal education,

explained a substantial amount of variation in the VAM estimates. In addition, the estimates for the year to year VAM resulted in systematic biases against schools with higher percentages of students receiving free and reduced lunch (Palardy & Peng, 2015). Papay (2011) also found differences in teacher-value added scores when students tests were measured from the fall of the first year to the fall of the second year rather than the typical spring to spring model. These results were replicated and confirmed by Attebery & Mangan (2020).

Because VAMs are frequently used for decisions such as teacher pay or even teacher retention, the impact of error rates in measuring teacher and school performance should be treated very seriously and interpreted with caution. Schochet and Chang (2013) define Type I error rate, or false positive, as finding a teacher who is actually average to be below average, and a Type II error rate, or false negative, as finding a teacher to be average when he or she is actually below average. Using a simulated data set and a system created to underestimate error rates, they found that Type I and Type II error rates for teacher-level analyses will be around 26% when using 3 years of data for estimation. For school-level analyses, however, they estimate the error scores will be 5-10 percentage points lower than in the teacher-level analyses due to a larger sample size. Their first model replaced unique student-level covariates with the teacher-level average, while the second model retained the student-level covariates. In their results, up to 26% of teachers in the bottom quintile using one model variation ranked higher than they did using an alternative model variation.

Traditional value-added models that ignore student mobility underestimate the importance of a school's contribution to overall variance. This occurs because the estimates obtained for between-school variance is smaller than the true between-school variance.

However, ignoring pupil movement will not cause serious errors in ranking schools against each other (Goldstein et al., 2007). Models honoring the cross-classified nature of the data compared to models containing purely nested structures produce only slight differences in fixed effects, but produce more substantial differences in random effects and their standard errors in their unconditional models (Grady & Beretvas, 2010). The difference in a cross-classified model and a standard two-level hierarchical model are the designation of effects as fixed or random. In a cross-classified model, Level 1 and Level 2 effects are regarded as random, where as in a standard two-level model the Level 1 group membership effect is fixed and the Level 2 effects are random (Raudenbush & Bryk, 2002).

Tekwe et al. (2004) compare VAMs under the HLM model with two newly proposed models, the layered mixed effects model (LMEM) and the simple fixed effects models (SFEM). The LMEM is the basis of the TVAAS. Unlike the HLM models, in which schools are assumed to be a random sample from a larger population of schools, the LMEM assumes schools to be taken from a fixed population of schools to be graded. Benefits of the LMEM are that it can be used with incomplete data and allows for multivariate analysis of several subjects simultaneously. The SFEM is appealing because it uses more simple calculations, making it more easily used and understood by principals and teachers without a significant statistics background. The SFEM calculates the school specific mean difference from the district wide mean, also treating schools as taken from a fixed population of schools to be graded Tekwe et al. (2004) found the SFEM and LMEM to be highly correlated. Because the SFEM is simpler, it is preferred.

Other VAM Applications

In addition to measuring teacher value-added scores for the purposes of accountability, VAMs have been recommended for other uses within education. Some research has suggested that raising teacher salaries will improve a school district's economic market. The cost will increase, but not enough to offset the reduced turnover. Teacher turnover can be costly to districts due to the increased cost of new teacher training. Monitoring the labor-supply response is an important area for future research (Rothstein, 2015; Koedel et al., 2015). Rather than deciding teacher pay or retention, Condie, Lefgren, and Sims (2014) suggest using VAMs to move teachers to contents in which they excel. Similarly, Glazerman et al. (2013) suggests moving high-value teachers to low-income areas. Doing so would keep the district's investment in each teacher while placing all teachers in an area where they can be successful.

Value-added models can also be useful in evaluation of programs. As part of the accountability process, teachers are evaluated yearly by the school principal or an assistant principal. Two separate studies have found a teacher's value-added scores and the teacher's yearly principal evaluation to have positive but weak correlations (Basileo & Toth, 2019; Harris et al, 2014). Little is still known about why this difference occurs and should be studied in future research. Finally, value-added models have recently been used to evaluate educational programs. Blau (2019) studied the impact of college courses on program satisfaction within a business school, while Brady et al. (2018) sought to improve teacher education programs by implanting new learning programs and an updated value-added assessment for teacher candidates.

References

Atteberry, A. & Mangan, D. (2020). The sensitivity of teacher value-added score to the use of fall or spring test scores. *Educational Researcher, 49(5)*, 335-349. https://doi.org/10.3102/0013189X20922993

Basileo, L. D. & Toth, M. (2019). A state level analysis of the Marzano Teacher Evaluation Model: Predicting teacher value-added measures with observation scores. *Practical Assessment, Research & Evaluation, 24(6).* https://doi.org/10.7275/cc5b-6j43

Blau, G. (2019). Integrating perceived added education value business administration core course items into scales and their relationships to degree program satisfaction and business school reputation influence. *Journal of Education and Learning, 8(4),* 1-7.

Brady, M. P., Miller, K., McCormick, J., & Heiser, L. A. (2018). A rational and manageable value-added model for teacher preparation programs. *Educational Policy*, *32*(5), 728-750.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, *40*, 76-92.

Glazerman, S., Protik, A., Teh, B. R., Bruch, J., & Max, J. (2013). Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment. NCEE 2014-4003. *National Center for Education Evaluation and Regional Assistance*.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(4),* 941-954. https://doi.org/10.1111/j.1467-985X.2007.00491.x

Grady, M. W., & Beretvas, S. N. (2010). Incorporating student mobility in achievement growth modeling: A cross-classified multiple membership growth curve model. *Multivariate Behavioral Research, 45*(3), 393-419. https://doi.org/10.1080/00273171.2010.483390

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal*, *51*(1), 73-112.

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration, 47(2),* 227-249. https://doi.org/10.1108/09578230910941066

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180-195.

Leckie, G. (2018). Avoiding bias when estimating the consistency and stability of value-added school effects. *Journal of Educational and Behavioral Statistics, 43(4)*, 440-468. https://doi.org/10.3102/1076998618755351

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational Behavioral Statistics, 29(1),* 67 – 101. https://doi.org/10.3102/10769986029001067

Nuttal, D. L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, *13*(7), 769-776.

Palardy, G. & Peng, L. (2015). The effect of summer on value-added assessments of teacher and school performance. *Educational Policy Analysis Archives, 23(92)*. http://dx.doi.org/10.14507/epaa.v23.1997

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48(1)*, 163-193. https://doi.org/10.3102/0002831210362589

Park, S. & Holloway, S. D. (2017). The effects of school-based parental involvement on academic achievement at the child and elementary school level: A longitudinal study. *The Journal of Educational Research, 110(1),* 1-16. https://doi.org/10.1080/00220671.2015.1016600

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic* Roy, M., & Giraldo-García, R. (2018). The role of parental involvement and social/emotional skills in academic achievement: Global perspectives. *School Community Journal*, *28*(2), 29-46.

Schochet, P. Z. & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38(2)*, 142-171. https://doi.org/10.3102%2F1076998611432174

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics 29*(1), 11-36. https://doi.org/10.3102%2F10769986029001011