THE IMPACT OF SUMMER LEARNING LOSS ON MEASURES OF SCHOOL PERFORMANCE

Andrew McEachin
(corresponding author)
RAND Corporation
Santa Monica, CA 90401
mceachin@rand.org

Allison Atteberry
School of Education
University of Colorado, Boulder
Boulder, CO 80309
Allison.Atteberry@colorado.edu

Abstract

State and federal accountability policies are predicated on the ability to estimate valid and reliable measures of school impacts on student learning. The typical spring-to-spring testing window potentially conflates the amount of learning that occurs during the school year with learning that occurs during the summer. We use a unique dataset to explore the potential for students' summer learning to bias school-level value-added models used in accountability policies and research on school quality. The results of this paper raise important questions about the design of performance-based education policies, as well as schools' role in the production of students' achievement.

doi:10.1162/EDFP_a_00213

1. INTRODUCTION

One of the most prominent debates in education policy today is how to design federal, state, and local policies that hold schools accountable for student outcomes. Such policies hinge upon the ability to estimate valid and reliable measures of school impacts on student learning that distinguish between schools' influence and the myriad of external factors that also contribute but are outside schools' purview. In response to this challenge, value-added models (VAMs) are being used to estimate schools' effects on students' achievement after conditioning on student and school characteristics that are assumed to be beyond the control of educators and administrators (Ladd and Walsh 2002; Todd and Wolpin 2003; Reardon and Raudenbush 2009; Guarino, Reckase, and Wooldridge 2015; Ehlert et al. 2016).

The validity of VAMs to produce unbiased measures of school performance rests on a number of assumptions, many of which have been explicated and probed in existing work (Todd and Wolpin 2003; Reardon and Raudenbush 2009; Chetty, Friedman, and Rockoff 2014a; Deming 2014; Guarino, Reckase, and Wooldridge 2015; Angrist et al. 2017). One important, but often ignored, assumption posits that the use of annual test scores, for tests usually administered each spring, measures schools' impact on student learning. Students' summer vacation, however, constitutes approximately a quarter of the days in the spring-to-spring testing window, which potentially conflates learning that occurs during the school year with learning that occurs during the summer.

Extant research on students' summer experiences suggests wide variation in time-use and learning, especially for low-income and minority students. White and middle-class children often exhibit learning gains over this time period, whereas minority and/or disadvantaged children experience losses (Alexander, Entwisle, and Olson 2001; Downey, von Hippel, and Broh 2004; Atteberry and McEachin 2016)—the negative impact of summer courses on lower socioeconomic students is often referred to as "summer setback" or "summer learning loss." Summer learning loss and nonrandom sorting of students across schools may complicate the measurement of schools' impact on student learning. However, even though a majority of schools are now held accountable for students' achievement growth from states' accountability policies under the federal Elementary and Secondary Education Act waivers (Polikoff et al. 2014), no research to date has examined the potential for summer setback to bias VAMs commonly used in research and school accountability policies.

In order to evaluate the impact of summer setback on school VAMs, we use a unique dataset from a southern state that contains both fall and spring test scores for students in grades 2 through 8. Specifically, we ask the following research questions:

RQ1. What is the magnitude and distribution of bias from students' differential summer learning in spring-to-spring school VAMs?

RQ2. How does the bias from students' differential summer learning affect the relative ranking of school quality as measured by schools' value added?

We find students' summer learning biases typical spring-to-spring VAMs, and that this bias negatively affects the relative standing of schools serving more disadvantaged students. The rest of the paper proceeds as follows: Section 2 reviews the relevant value-added modeling and summer learning literature; section 3 discusses the unique dataset used to answer our research questions; section 4 describes our methods; and sections 5 and 6 present the results and concluding remarks.

2. LITERATURE REVIEW

Policies that hold teachers and schools accountable for their students' outcomes have been implemented for two main reasons: to solve the principal-agent problem and to address market failures due to information asymmetry (Holmstrom and Milgrom 1991; Prendergast 1999; Baker 2000; Figlio and Lucas 2004; Figlio and Kenny 2009). The former assumes the use of performance incentives will better align educators' behaviors with local, state, or federal standards (Holmstrom and Milgrom 1991; Smith and O'Day 1991; Prendergast 1999). The latter aims to infuse the educational marketplace with information about teacher and school quality (Rothstein, Jacobsen, and Wilder 2008; Figlio and Loeb 2011; Charbonneau and Van Ryzin 2012). In both cases, performance is generally defined in terms of student achievement on standardized tests, which presumes that test scores, despite being narrow in scope, predict students' long-term success (Chetty et al. 2011; Chetty, Friedman, and Rockoff 2014b).

Accountability policies need valid measures of teacher and school performance to elicit optimal behavior from educators. If these measures are too noisy, too rigid, or biased by factors unrelated to schooling activities, incentives to align behaviors with expectation break down and unintended consequences may emerge. Previous research has examined the potential for bias in VAMs, as well as the sensitivity of VAMs to model specification, measurement error, and year-to-year stability (McCaffrey et al. 2009; Papay 2011; Ehlert et al. 2013; Goldhaber and Hansen 2013; Chetty et al. 2014a; Angrist et al. 2017). Few studies, however, have evaluated the relationship between students' summer learning and school quality or VAMs (Downey, von Hippel, and Hughes 2008; Papay 2011; Palardy and Peng 2015; Gershenson and Hayes 2016).

Studies using both local and national datasets have found that both socioeconomic and race gaps in reading skills grew at faster rates during the summer (Heyns 1978; Entwisle and Alexander 1992, 1994; Burkham et al. 2004; Downey, von Hippel, and Broh 2004; Atteberry and McEachin 2016). It is unclear what causes students of different backgrounds to have different summer experiences, though research has suggested that income differences could be related to students' opportunities to practice academic skills and learn over summer (Heyns 1978; Cooper et al. 1996; Downey et al. 2004). For example, Gershenson (2013) found that low-income students watch two more hours of television per day during the summer than students from wealthier backgrounds. Burkham et al. (2004) and Gershenson and Hayes (2016), however, found that even a rich set of student and family observables explains a very small share of variation in students' summer math and reading learning.

Little attention has been paid to the intersection between summer learning loss and VAMs, partially due to the ubiquitous spring-to-spring test timeline in school accountability policies and large education datasets. To date, we know of only four papers that investigate the role that summers play in teacher and school accountability (Downey, von Hippel, and Hughes 2008; Papay 2011; Palardy and Peng 2015; Gershenson and Hayes 2016). Although Papay (2011) did not directly investigate the role of summer learning loss in teacher VAMs, he found Spearman rank correlations between teachers' spring–spring and fall–spring math and reading value added of 0.66 and 0.71, respectively. Gershenson and Hayes (2016), using data from the Early Childhood Longitudinal Study of 1998–99, not only found similar rank correlations but also found the inclusion of rich detail about the students' summer activities (including summer school attendance) and parents' backgrounds in spring-to-spring VAMs did not improve the correlations. Lastly, Downey, von Hippel, and Hughes (2008) and Palardy and Peng (2015), also using Early Childhood Longitudinal Study data, estimated random-effect growth models and found that schools serving larger shares of low-income students were more likely to be in the bottom of the performance distribution when school performance measures were based on annual rather than school-year learning.

The four studies suggest that ignoring the summer period will produce a systematic bias in VAMs that may also disproportionately affect schools serving larger shares of minority and low-income students under traditional accountability regimes. The studies also raise unanswered questions. First, Papay (2011) and Gershenson and Hayes (2016) do not investigate the relationship between the discordance in VAMs from different test timelines and student/school demographics. Second, the four studies also have important data limitations, either relying on a few years of data within one urban district (Papay 2011), or one summer between kindergarten and first grade for one cohort of nationally representative students (Downey, von Hippel, and Hughes 2008; Palardy and Peng 2015; Gershenson and Hayes 2016).

The results of our paper address gaps in the interrelated accountability, value-added modeling, and summer learning literatures in three ways. First, we utilize a state-wide panel of student achievement data from grades 2 through 8 over a five-year period. Instead of relying on one summer between kindergarten and first grade, the grade span used in this study is more representative of grades typically included in high-stakes accountability. Second, we are the first to evaluate the impact of summer setback on traditional school VAMs that are becoming popular in state and federal accountability policies. Lastly, we not only examine whether summer setback leads to potential misclassifications in schools' rank-ordering of math and reading value added, but also the types of schools most affected by this phenomenon.

3. DATA

We use Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) assessment data for our analysis. A computer adaptive test, MAP assesses student performance in math, reading, and language arts, and science and is administered to students in all fifty states in the United States, although the share of MAP examinees in each state varies from less than 5 percent to nearly all students. To ensure that MAP scores provide a valid measure of student achievement, NWEA aligns MAP items with state standards (including the new Common Core State Standards). The MAP assessment is scored using a vertical and equal-interval scale, which NWEA refers to as the Rasch Unit (RIT) scale. The vertical scale allows comparisons of student learning across grades and over time, and the equal-interval scale ensures that unit increases represent equal learning gains across the entire distribution.[1] In sum, the MAP assessment

[1] We do not standardize students' reading or math MAP scores because the scores are on a vertical and equal-interval scale, and are normally distributed.

has appropriate measurement properties for a study of school VAMs and summer setback.

We utilize data from a southern state that administered the MAP assessment in fall and spring semesters for all students in grades 3 through 8 for the 2007–08 through 2010–11 school years. In this context, the MAP was used as a formative assessment to provide teachers and principals information about their students' start- and end-of-year achievement. We discuss the limitations of using low-stakes assessment data in section 5. Our data include student- and school-level files that are longitudinally matched over time.[2] The student-level file includes basic demographic information, such as students' race and gender, their math and reading scores, the measurement error associated with their math and reading scores, grade of enrollment, the date of test administration, and fall and spring school identifiers. Notably, the student-level file does not include indicators for whether the student is an English Language Learner, belongs to the federal Free and Reduced Price Lunch (FRPL) program, or participates in special education. It is unlikely the omission of these variables will bias our results. As noted above, Gershenson and Hayes (2016) find that a rich set of covariates, including detailed socioeconomic data and students' and parents' summer activities, explain only 3 to 5 percent of the variation in students' summer learning.

The school-level data file is provided by the Common Core of Data through NWEA, and includes a typical set of school-level characteristics, including the percent of FRPL-eligible students within a school. The student- and school-level descriptives for the 2010–11 school year are provided in table 1.

Following Quinn (2015), we use school calendar information to separate the portion of students' learning that occurs during the school year from the summer period. As discussed in more detail in Appendix B, we project students' learning to the first day of the fall semester and the last day of the spring semester. This projection removes the instructional time that typically occurs between when students take the spring MAP assessment and the end of the school year, as well as the time between the start of the school year and when students take the fall MAP assessment. We use students' projected fall and spring achievement as dependent and independent variables throughout our analysis.

We use VAMs to estimate three-year average school effects on students' math and reading achievement. For the spring-to-spring timeline, we use students' prior spring test score as the control for prior achievement. Because our data window starts with the 2007–08 school year, the first year we can generate schools' value added using the spring-to-spring test timeline is 2008–09. The use of a lagged spring achievement variable further restricts the sample to students in at least third grade. Our analytic sample for all models includes students in grades 3 through 8 during the 2008–09 to 2010–11 school years with approximately 45,000 students per grade per year.

4. METHODS

Although students' true education production function is unknown to the researcher, extant research suggests that a simple dynamic ordinary least squares (DOLS)

[2] The data received from NWEA have been scrubbed of any identifying variables (or IDs) for students, schools, and districts.

Table 1. Student Demographics for the 2010–11 School Year

	Mean	SD	*N*
Student demographics			
Spring math MAP achievement	224.22	19.67	222,765
Spring reading MAP achievement	214.89	17.04	219,665
Fall math MAP achievement	212.08	20.16	222,765
Fall reading MAP achievement	206.07	18.75	222,765
Summer math Map loss	−4.21	9.50	163,495
Summer reading Map loss	−2.50	11.25	161,825
Lagged spring math MAP achievement	216.33	20.48	222,765
Lagged spring reading MAP achievement	208.86	18.31	222,765
White student	0.530		222,765
Black student	0.363		222,765
Hispanic student	0.059		222,765
Mobile student (between school years)	0.079		222,765
Mobile student (within school years)	0.029		222,765
3rd grade	0.171		222,765
4th grade	0.182		222,765
5th grade	0.183		222,765
6th grade	0.160		222,765
7th grade	0.157		222,765
8th grade	0.148		222,765
% Hispanic in school	0.063		222,765
% Black in school	0.357		222,765
% White in school	0.527		222,765
% FRPL in school	0.577		222,765
Urban school	0.169		222,765
Suburban school	0.265		222,765
Town school	0.145		222,765
Rural school	0.421		222,765
School enrollment	673.17	254.51	222,765
School demographics			
% Hispanic in school	0.064		766
% Black in school	0.407		766
% White in school	0.490		766
% FRPL in school	0.625		766
Urban school	0.173		767
Suburban school	0.216		767
Town school	0.155		767
Rural school	0.455		767
School enrollment	561.94	234.22	767

model—which regresses current achievement on prior achievement, school fixed-effects, and vectors of student and school control variables—produces unbiased estimates of schools' value added (Deming 2014; Guarino, Reckase, and Wooldridge 2015). The DOLS specification assumes that effects of current inputs are captured by students' current observable characteristics and school inputs, prior inputs are captured by students' lagged achievement (including her fixed academic endowment), and the effect of these inputs on current achievement decays at a constant geometric rate.

We start with the standard DOLS specification:

$$Y_{ist} = \theta_1 Y_{is(t-1)} + \theta_2 \tilde{Y}_{is(t-1)} + \boldsymbol{\beta X}_{ist} + \boldsymbol{\zeta Z}_{st} + \boldsymbol{\lambda \delta}_s + \boldsymbol{\alpha}_t + \varepsilon_{ist}, \tag{1}$$

where Y_{ist} is modeled as a linear additively separable function of the spring achievement in the prior school year $t - 1$ in the same and off-subject $Y_{is(t-1)}$ and $\tilde{Y}_{is(t-1)}$; a vector of student demographic characteristics $\boldsymbol{X}_{ist}$, including race, a mobility indicator for whether the student made a nonstructural school move, an indicator for whether students changed schools within the school year, and indicators for students' grade-level; school-level aggregates of the student demographic characteristics $\boldsymbol{Z}_{st}$, as well as additional controls for the percent of FRPL students in the school and the natural log of enrollment; a vector of school indicator variables $\boldsymbol{\delta}_s$, which take a one for the school to which the student is exposed in the given year and a zero otherwise; year fixed-effects $\boldsymbol{\alpha}_t$; and an idiosyncratic student-level error term ε_{ist}.[3] The key parameters of interest are $\boldsymbol{\lambda}$, which capture the average conditional achievement of a school's students over the three-year panel.[4,5]

Given that most state-wide testing systems use a spring-to-spring test timeline, researchers and policy makers have to use students' prior spring achievement, $Y^{Spring}_{is(t-1)}$, to capture the effect of prior inputs on current achievement. Roughly three months of the time between $Y^{Spring}_{is(t-1)}$ and Y^{Spring}_{ist} occur outside of the school year, potentially conflating school-year learning and summer learning loss in $\boldsymbol{\lambda}$. If instead we wanted to measure the impact of schools on school-year learning, we would want to condition on students' achievement on the first day of school Y^{Fall}_{ist} instead of $Y^{Spring}_{is(t-1)}$. For example, we define students' fall achievement as $Y^{Fall}_{ist} = SL_{ist} + Y^{Spring}_{is(t-1)}$, where SL_{ist} captures students' summer learning. Only when $SL_{ist} = 0$ will students' fall achievement equal their prior spring achievement. If $SL_{ist} \neq 0$ students' summer learning is part of the error term in equation 1:

$$Y^{Spring}_{ist} = \theta_1 Y^{Spring}_{is(t-1)} + \theta_2 \tilde{Y}^{Spring}_{is(t-1)} + \boldsymbol{\beta X}_{ist} + \boldsymbol{\zeta Z}_{st} + \boldsymbol{\lambda \delta}_s + \boldsymbol{\alpha}_t + (\theta_1 SL_{ist} + \theta_2 \widetilde{SL}_{ist} + \eta_{ist}). \tag{2}$$

[3] We use t to index the typical school year. For example, if we define $t \equiv$ 2014–15 school year, $t - 1$ would end on the last day of school in the spring 2014 and t would start the first day of summer 2014 and run through the last day of school in spring 2015.

[4] Although not shown, we also estimate models with school-by-year fixed effects capturing the year-to-year impact on students' math and reading achievement. The results of the school-by-year model are qualitatively similar to the results presented in this paper. The school-by-year specification allows us to not only tease out important differences in schools' value added by changing the test timeline but it also allows us to see if year-to-year variation in value added for a given test timeline is related to school demographics. For example, we looked at whether the year-to-year fluctuation in spring-to-spring value added generates results similar to those presented in this paper when comparing spring-to-spring with fall-to-spring value added. We find that the year-to-year fluctuation in spring-to-spring value added is not correlated with student demographics. This suggests that our results in this paper capture true differences in estimates of school quality and not random noise. We prefer the specification in equation 1 that uses multiple years of data to generate a single value-added estimate.

[5] As explained in more detail in Appendix A, the λ are estimated using a sum-to-zero constraint to ease interpretation, centering λ on the state grand mean. Schools with a positive λ have value added above the state average, and vice versa for schools with a negative λ. We run the model separately for math and reading. The inclusion of student characteristics (including prior achievement) and school fixed effects accounts for the sorting of student to schools by these observable characteristics, generating school value added net of these background characteristics.

If students' summer learning, SL_{ist} and $\widetilde{SL}_{ist}$, are correlated with the independent variables in equation 2, then schools' value-added estimates from the spring-to-spring test timeline will be different than those generated from a fall-to-spring test timeline.[6] Whether explicit or not, researchers and practitioners who use equation 2 to hold schools (or teachers) accountable for student achievement assume either that schools are responsible for students' summer learning or students' summer learning does not predict current spring achievement conditional on observable student and school characteristics.

Extant research, and the results of our analysis, show that students' summer learning does vary across schools in systematic ways although the vast majority of the variation in students' summer learning is not explained by between-school differences (i.e., differences in practices or resources). We assume for the moment that students' summer learning is the main source of bias of interest in equation 1, and any bias left over in a fall-to-spring VAM is also common to a spring-to-spring VAM. For example, if students sort to schools based on factors unobserved to the researcher, this bias would likely equally affect VAMs from either timeline. If this is true, our results capture the relative bias caused by summer learning loss between two models.[7] We estimate the bias in schools' spring-to-spring math and reading value added from students' summer learning in two ways.

First, we generate spring-to-spring and fall-to-spring math and reading value-added estimates from equation 1, using the same vector of student and school covariates in each model. If, on average, schools' spring-to-spring value added is an unbiased estimate of schools fall-to-spring value added, then the coefficient of a simple OLS regression of schools' fall-to-spring value added on their spring-to-spring value added will be statistically indistinguishable from 1 (Kane and Staiger 2008; Chetty, Friedman, and Rockoff 2014a; Deming 2014).

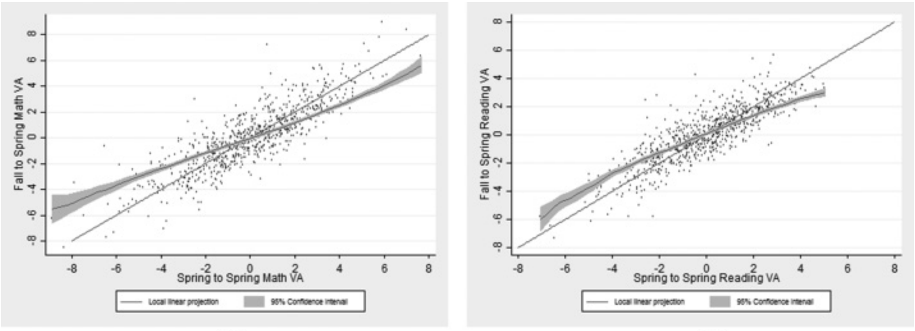
Second, we estimate the unit specific bias in schools' math and reading value added using the standard omitted variable bias framework (c.f. a standard treatment in Greene 2003 or Angrist and Pischke 2009 and as it applies to teacher VAMs in Goldhaber and Chaplin 2015). Specifically, we estimate the bias in schools' spring-to-spring value added as

$$Bias\left(\hat{\boldsymbol{\lambda}}\right) = E\left[\hat{\boldsymbol{\lambda}}\right] - \boldsymbol{\lambda} = \frac{Cov\left(\boldsymbol{\delta}_s^*, SL_{ist}^*\right)}{Var\left(\boldsymbol{\delta}_s^*\right)}\theta_1^{Summer} + \frac{Cov\left(\boldsymbol{\delta}_s^*, \widetilde{SL}_{ist}^*\right)}{Var\left(\boldsymbol{\delta}_s^*\right)}\theta_2^{Summer}, \tag{3}$$

where $\boldsymbol{\delta}_s^*$, SL_{ist}^*, and $\widetilde{SL}_{ist}^*$ are residualized school indicators and measures of students' math and reading summer learning, respectively, using the student and school covariates from equation 1. The terms θ_1^{Summer} and θ_2^{Summer} are coefficients for SL_{ist} and $\widetilde{SL}_{ist}$ if they were included in equation 2.

[6] Our analysis in this paper ignores other sources of noise likely present in estimating and evaluating VAMs (i.e., measurement error, sampling variation). We assume these are common to models that do and do not include students' summer learning, and therefore do not change the qualitative interpretation of our findings.

[7] There are recent strong claims about the ability of a simple VAM to estimate the causal impact of teachers and schools on students' achievement (e.g., Chetty, Friedman, and Rockoff 2014a; Deming 2014). Recent work by Angrist et al. (2017), however, suggests that a small amount of bias does exist in the traditional VAM.



(a) (b)

Figure 1. Scatter Plot of Schools' Fall-to-Spring Value Add on Spring-to-Spring Value Added.

The terms $\frac{Cov(\delta_s^*, SL_{ist}^*)}{Var(\delta_s^*)} = \lambda^{Summer}$ and $\frac{Cov(\delta_s^*, \widetilde{SL}_{ist}^*)}{Var(\delta_s^*)} = \tilde{\lambda}^{Summer}$ in equation 3 capture the average amount of summer learning within school *s* conditional on students' prior achievement and student and school covariates. We estimate λ^{Summer} and $\tilde{\lambda}^{Summer}$ from separate auxiliary DOLS regressions of equation 1 with students' summer learning as the dependent variable. We use separate joint F-tests for the restrictions that $\lambda^{Summer} = 0$ and $\tilde{\lambda}^{Summer} = 0$ to test whether students' summer learning is equally distributed across schools. A failure to reject these F-tests suggests that students' summer learning does not vary across schools.

We could also evaluate the impact of students' summer learning loss on school VAMs by taking the difference of schools' spring-to-spring and fall-to-spring value added, and evaluating how this difference is related to school characteristics. In supplementary analyses we replicate our results using the difference of schools' value added instead of equation 3; the results are qualitatively similar and are available upon request.

The bias in school value added from students' summer learning is particularly policy-relevant if the correlation between the estimated bias and the share of traditionally under-served students within a school is negative, penalizing schools for educating students from disadvantaged backgrounds.

5. RESULTS

RQ1: What is the Magnitude and Distribution of Bias from Students' Differential Summer Learning in Spring-to-Spring School Value-Added Models?

We start our analysis with visual inspection of the relationship between these two measures of schools' math and reading value added in panels A and B of figure 1. The coefficients for the VAMs can be found in table C.1 in Appendix C. If schools' spring-to-spring math and reading value added are an unbiased estimate of the true effect of schools on students' achievement, then points will be tightly clustered around the 45° line, with small deviations scattered randomly throughout the joint distribution. As the bias from students' summer learning increases, the deviations between the points in the scatter plot and the 45° line will increase and the slope of a line through the scatter plot will deviate from the 45° line. Even though the points in figure 1 are roughly scattered around the 45° line, the relationship between $\hat{\lambda}^{Fall}$ and $\hat{\lambda}^{Spring}$ does suggest a small systematic bias. Ninety-five percent confidence intervals for local polynomial

Table 2. Regression of Schools' Fall-to-Spring Value Added on Spring-to-Spring Value Added

	OLS	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$
Schools' spring-to-spring **math** value added	0.854***	0.919***	0.871***	0.822***	0.823***	0.875***
	(0.021)	(0.031)	(0.020)	(0.021)	(0.029)	(0.039)
Adjusted R^2	0.725					
# of Schools	774	774	774	774	774	774
p-value ($\phi = 1$)	0.000	0.010	0.000	0.000	0.000	0.001
Schools' spring-to-spring **reading** value added	0.833***	0.905***	0.848***	0.820***	0.780***	0.767***
	(0.022)	(0.028)	(0.024)	(0.020)	(0.024)	(0.033)
Adjusted R^2	0.727					
# of Schools	774	774	774	774	774	774
p-value ($\phi = 1$)	0.000	0.001	0.000	0.000	0.000	0.000

***$p < 0.001$.

regressions through the scatter plots in figure 7 show that, on average, points do not fall on the 45° line—although the average deviation appears modest.

Results of OLS and quantile regressions of schools' $\hat{\lambda}^{Fall}$ on $\hat{\lambda}^{Spring}$ are presented in table 2.[8] For both math and reading, we reject the hypothesis $\hat{\phi}_1 = 1$ in the OLS specification and across all various quantiles of $\hat{\lambda}^{Fall}$. The projection bias, $1 - \hat{\phi}$, from using a spring-to-spring test timeline to estimate equation 1, ranges between 8 and 24 percent depending on the subject and placement of the school in the $\hat{\lambda}^{Fall}$ distribution. The projection bias in table 2 can be characterized by a simple thought experiment. Imagine a policy that moved students from a low to high performing school, defined by a one standard deviation increase in math $\hat{\lambda}^{Spring}$ (2.44 RIT points). As a result of that policy change, we would expect students' within-year learning to increase only $.85 * \sigma_{\hat{\lambda}^{Spring}}$. The projection bias in table 2 is roughly one-third the size of the bias from a poorly specified model (e.g., a model that does not condition on students' prior achievement; Deming 2014). The results in figure 1 and table 2 indicate an average bias in schools' spring-to-spring value added exists, but do not describe the distribution of the bias.

Next, we focus on the results of the school-specific bias in schools' spring-to-spring value added from students' differential summer learning. As shown in the first row of panel A in table 3, we reject the null hypothesis that $\hat{\lambda}^{Summer} = 0$ for both math and reading. Furthermore, we construct 95 percent confidence intervals for $\hat{\lambda}^{Summer}$ to evaluate how many schools have negative, zero, and positive estimates of conditional math and reading summer learning loss. Panel A indicates that 21 and 26 percent, respectively, of the schools have negative estimates of aggregate conditional math and reading summer learning, and 28 and 32 percent, respectively, have positive estimates of aggregate conditional math and reading summer learning. If the second condition for students' summer learning to bias schools' spring-to-spring value added is met, approximately 50 percent of the schools in our sample would have a different math or reading value added in a fall-to-spring timeline than a spring-to-spring timeline because of students' differential summer learning.

[8] We test whether our results are sensitive to the fact that both $\hat{\lambda}^{Fall}$ and $\hat{\lambda}^{Spring}$ are estimated. First we run this, and all other analyses, using empirical Bayes shrunk values of $\hat{\lambda}^{Fall}$ and $\hat{\lambda}^{Spring}$. Second, we estimate $\hat{\lambda}^{Fall} = \phi_0 + \phi_1 \hat{\lambda}^{Spring} + \varepsilon_s$ weighting by the precision of $\hat{\lambda}^{Fall}$, $\frac{1}{se(\hat{\lambda}^{Fall})}$. In both cases our results are nearly identical to those presented in this paper.

Table 3. Analysis of Potential Bias in Schools' Math and Reading Value Added from Students' Summer Learning

	Math	Reading
Panel A: $\frac{Cov(\delta_s^*, SL_{ist}^*)}{\delta_s^*}$		
p-value $\hat{\lambda}_1^{Summer} = \hat{\lambda}_2^{Summer} = \cdots = \hat{\lambda}_S^{Summer}$	0.000	0.000
Share of $\hat{\lambda}^{Summer} > 0$	0.213	0.263
Share of $\hat{\lambda}^{Summer} = 0$	0.512	0.419
Share of $\hat{\lambda}^{Summer} < 0$	0.275	0.318
Panel B: θ_1^{Summer} and $\tilde{\theta}_2^{Summer}$		
$\hat{\theta}_1^{Summer}$	0.243^{***} (0.004)	0.235^{***} (0.003)
$\widehat{\tilde{\theta}}_2^{Summer}$	0.153^{***} (0.002)	0.116^{***} (0.003)
Panel C: Distribution of Bias in Schools' Value Added		
Standard deviation of unit-specific bias	0.619	0.547
Standard deviation of spring-to-spring value added	2.448	2.062
Pearson correlation (bias, % FRPL)	−0.616	−0.614
Pearson correlation (bias, % Minority)	−0.408	−0.421

Note: The regressions in panels A and B include 641,099 student/year observations. $^{***}p < 0.001$.

As shown in panel B of table 3, students' same-subject and off-subject summer learning has a positive, statistically significant relationship with students' current spring achievement. The results in panels A and B indicate that, even conditional on students' prior spring achievement and student and school covariates, students' summer learning meets the necessary and sufficient criteria to bias in schools' spring-to-spring math and reading value added. Finally, the results in panel C show that standard deviation of the bias from students' summer learning is roughly 25 percent of the standard deviation of schools' spring-to-spring value added, and this bias is strongly negatively correlated with the percent of FRPL students ($r = -0.61$) and minority students ($r = -0.41$) in a school.

We present graphical representations of the bias in panels A and B of figure 2. Specifically, we plot the kernel density of the bias in schools' math and reading spring-to-spring value added by tertiles of the share of FRPL students in a school. These figures show the difference in mean school-specific bias between the poorest tertile and the wealthiest tertile is approximately 1 point on the MAP RIT scale, roughly equal to 2 standard deviations of the bias in spring-to-spring value added or 5 percent of a standard deviation in students' math and reading MAP achievement. In the next section, we further explore which schools are potentially most affected by the bias.

RQ2: How Does the Bias from Students' Differential Summer Learning Affect the Relative Ranking of School Quality as Measured by Schools' Value Added?

To answer the second research question, we start with a simple examination of the Spearman rank-order correlation among schools' spring-to-spring and fall-to-spring value added, and aggregate student demographics, as shown in table 4. The correlation

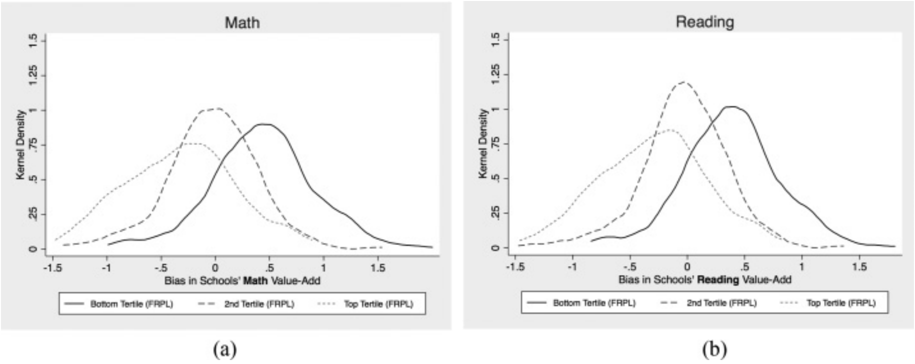


Figure 2. Kernel Density of the Bias in Schools' Math and Reading Value Added.

Table 4. Correlation of Schools' Math and Reading Value Added and Demographics

	Math $\hat{\lambda}^{Spring}$	Math $\hat{\lambda}^{Fall}$	Reading $\hat{\lambda}^{Spring}$	Reading $\hat{\lambda}^{Fall}$	Percent FRPL	Percent Minority
Math $\hat{\lambda}^{Spring}$	1.00					
Math $\hat{\lambda}^{Fall}$	0.85	1.00				
Reading $\hat{\lambda}^{Spring}$	0.61	0.40	1.00			
Reading $\hat{\lambda}^{Fall}$	0.43	0.49	0.85	1.00		
Percent FRPL	−0.51	−0.22	−0.60	−0.34	1.00	
Percent minority	−0.50	−0.22	−0.50	−0.30	0.70	1.00

of schools' math and reading rankings between $\hat{\lambda}^{Spring}$ and $\hat{\lambda}^{Fall}$ timeline is relatively stable ($r = 0.85$). Similar to extant research in teacher value added (Papay 2011), the cross-subject rankings, however, are lower, ranging from $r = 0.4$ to 0.6. Although our school-level within-subject correlations of $\hat{\lambda}^{Spring}$ and $\hat{\lambda}^{Fall}$ are stronger than the year-to-year correlations of teacher value added (McCaffrey et al. 2009), it is still possible for the discordance between test-timelines to have important implications for school accountability systems.

In the last two rows of table 4, we show for both math and reading that the spring-to-spring value added have strong negative correlations with the percent of FRPL students in a school: $r_{math} = -0.51$ and $r_{reading} = -0.60$, and the percent of minority students in a school $r_{math} = -0.50$ and $r_{reading} = -0.56$. Schools' math and reading fall-to-spring value added have weaker negative correlations with the percent of FRPL students: $r_{math} = -0.22$ and $r_{reading} = -0.34$, and the percent of minority students in a school $r_{math} = -0.22$ and $r_{reading} = -0.29$. These differences are statistically significant ($p \leq 0.01$) (Meng, Rosenthal, and Rubin 1992) and suggest that switching to a fall-to-spring test timeline qualitatively changes the types of schools possibly identified for rewards and/or sanctions under an accountability system.

In panel A of tables 5 and 6 we present transition matrices across quintiles of schools' math and reading spring-to-spring and fall-to-spring value added. Similar to Gershenson and Hayes (2016), and to a lesser degree Papay (2011), we find nontrivial differences in the quintile ranking between the two test timelines. We focus on the math results in table 5 for the sake of brevity, but the results are similar for reading.

Table 5. Transition Matrix for Schools' Math Fall-to-Spring ($\hat{\lambda}^{Fall}$) and Spring-to-Spring ($\hat{\lambda}^{Spring}$) Value Added

Panel A: Full Sample

	(Bottom)				(Top)	
	Q_1 $\hat{\lambda}^{Fall}$	Q_2 $\hat{\lambda}^{Fall}$	Q_3 $\hat{\lambda}^{Fall}$	Q_4 $\hat{\lambda}^{Fall}$	Q_5 $\hat{\lambda}^{Fall}$	Total
(Bottom) Q_1 $\hat{\lambda}^{Spring}$	107	33	12	2	1	155
Q_2 $\hat{\lambda}^{Spring}$	40	67	37	10	1	155
Q_3 $\hat{\lambda}^{Spring}$	7	42	62	40	4	155
Q_4 $\hat{\lambda}^{Spring}$	1	12	36	72	34	155
(Top) Q_5 $\hat{\lambda}^{Spring}$	0	1	8	31	114	154
Total	155	155	155	155	154	

Panel B: Bottom Quintile FRPL (Least Poor)

	(Bottom)				(Top)	
	Q_1 $\hat{\lambda}^{Fall}$	Q_2 $\hat{\lambda}^{Fall}$	Q_3 $\hat{\lambda}^{Fall}$	Q_4 $\hat{\lambda}^{Fall}$	Q_5 $\hat{\lambda}^{Fall}$	Total
(Bottom) Q_1 $\hat{\lambda}^{Spring}$	7	1	1	0	0	9
Q_2 $\hat{\lambda}^{Spring}$	9	1	0	0	0	10
Q_3 $\hat{\lambda}^{Spring}$	2	9	10	2	0	23
Q_4 $\hat{\lambda}^{Spring}$	1	10	13	19	2	45
(Top) Q_5 $\hat{\lambda}^{Spring}$	0	1	5	15	47	68
Total	19	22	29	36	49	

Panel C: Top Quintile FRPL (Most Poor)

	(Bottom)				(Top)	
	Q_1 $\hat{\lambda}^{Fall}$	Q_2 $\hat{\lambda}^{Fall}$	Q_3 $\hat{\lambda}^{Fall}$	Q_4 $\hat{\lambda}^{Fall}$	$Q_5\hat{\lambda}^{Fall}$	Total
(Bottom) Q_1 $\hat{\lambda}^{Spring}$	39	20	8	1	1	69
Q_2 $\hat{\lambda}^{Spring}$	2	8	13	5	1	29
Q_3 $\hat{\lambda}^{Spring}$	0	2	10	17	4	33
Q_4 $\hat{\lambda}^{Spring}$	0	0	0	5	13	18
(Top) Q_5 $\hat{\lambda}^{Spring}$	0	0	0	0	5	5
Total	41	30	31	28	24	

Notes: FRPL = free and reduced price lunch; Q = quintile.

For example, of 155 schools in the bottom quintile in $\hat{\lambda}^{Spring}$, 48 of them (31 percent) are in higher quintiles of $\hat{\lambda}^{Fall}$, including 24 in at least the third quintile. Similar patterns emerge when comparing the share of schools in the bottom quintile of $\hat{\lambda}^{Fall}$. Furthermore, there is similar movement for schools in the fourth quintile of either $\hat{\lambda}^{Spring}$ or $\hat{\lambda}^{Fall}$. Approximately 50 of the 155 schools in the fourth quintile in either test timeline are in a lower quintile in the opposite timeline, while 30 are in the top quintile. Movement among the quintiles between the two test timelines is especially important if it is related to school demographics in systematic ways.

We evaluate this possibility in tables 5 and 6 by reporting the transition matrices for schools in the bottom quintile of percent FRPL (least poor) in panel B and the top quintile of percent FRPL (most poor) in panel C. One immediate, and unsurprising, pattern emerges: Schools with lower shares of FRPL students are clustered in the top quintiles of value added in both timelines, and schools with larger shares of FRPL students are clustered in the bottom quintiles of value added in both timelines. As suggested by the negative relationship between equation 3 and school poverty, the relationship between school poverty and school performance is stronger in the spring-to-spring time than

Table 6. Transition Matrix for Schools' Reading Fall-to-Spring ($\hat{\lambda}^{Fall}$) and Spring-to-Spring ($\hat{\lambda}^{Spring}$) Value Added

Panel A: Full Sample

	(Bottom)				(Top)	
	Q_1 $\hat{\lambda}^{Fall}$	Q_2 $\hat{\lambda}^{Fall}$	Q_3 $\hat{\lambda}^{Fall}$	Q_4 $\hat{\lambda}^{Fall}$	Q_5 $\hat{\lambda}^{Fall}$	Total
(Bottom) Q_1 $\hat{\lambda}^{Spring}$	110	34	7	2	2	155
Q_2 $\hat{\lambda}^{Spring}$	38	59	47	9	2	155
Q_3 $\hat{\lambda}^{Spring}$	7	48	62	31	7	155
Q_4 $\hat{\lambda}^{Spring}$	0	12	36	69	38	155
(Top) Q_5 $\hat{\lambda}^{Spring}$	0	2	3	44	105	154
Total	155	155	155	155	154	

Panel B: Bottom Quintile FRPL (Least Poor)

	(Bottom)				(Top)	
	Q_1 $\hat{\lambda}^{Fall}$	Q_2 $\hat{\lambda}^{Fall}$	Q_3 $\hat{\lambda}^{Fall}$	Q_4 $\hat{\lambda}^{Fall}$	$Q_5\hat{\lambda}^{Fall}$	Total
(Bottom) Q_1 $\hat{\lambda}^{Spring}$	4	1	0	0	0	5
Q_2 $\hat{\lambda}^{Spring}$	3	3	2	0	0	8
Q_3 $\hat{\lambda}^{Spring}$	3	8	11	2	0	24
Q_4 $\hat{\lambda}^{Spring}$	0	8	11	16	5	40
(Top) Q_5 $\hat{\lambda}^{Spring}$	0	1	2	29	46	78
Total	10	21	26	47	51	

Panel C: Top Quintile FRPL (Most Poor)

	(Bottom)				(Top)	
	Q_1 $\hat{\lambda}^{Fall}$	Q_2 $\hat{\lambda}^{Fall}$	Q_3 $\hat{\lambda}^{Fall}$	Q_4 $\hat{\lambda}^{Fall}$	$Q_5\hat{\lambda}^{Fall}$	Total
(Bottom) Q_1 $\hat{\lambda}^{Spring}$	56	19	5	1	1	82
Q_2 $\hat{\lambda}^{Spring}$	4	11	14	6	2	37
Q_3 $\hat{\lambda}^{Spring}$	0	0	5	8	6	19
Q_4 $\hat{\lambda}^{Spring}$	0	0	1	1	6	8
(Top) Q_5 $\hat{\lambda}^{Spring}$	0	0	0	0	8	8
Total	60	30	25	16	23	

the fall-to-spring timeline. For example, schools in panel B of table 5 are 2.2 times *more* likely to be in the bottom two quintiles of $\hat{\lambda}^{Fall}$ compared with $\hat{\lambda}^{Spring}$, whereas schools in panel C are 1.4 times more likely to be in the bottom two quintiles of $\hat{\lambda}^{Spring}$ compared with $\hat{\lambda}^{Fall}$. The same holds at the top end of the distribution.

Another way to think about the difference in schools' relative performance between the two test timelines is to examine the share of schools in a given performance quintile in panel B or C. If the relationship between school poverty and quality (as measured by value added) was independent, each row and column total would sum to approximately 31. The row totals for $\hat{\lambda}^{Spring}$ are heavily skewed toward the top quintile in panel B and toward the bottom quintile in panel C. The column totals are more equally distributed for $\hat{\lambda}^{Fall}$, ranging from 19 to 49 for panel B and 24 to 41 in panel C.

In table 7 we take a final look at how the bias in schools' spring-to-spring value added from students' summer learning affects the inferences made on school quality. Specifically, we look at the share of schools in a higher, equal, or lower quintile of spring-to-spring value-added compared with fall-to-spring across quintiles of percent of FRPL students in a school. Table 7 shows that inferences about school quality from a spring-to-spring test timeline favor schools serving lower shares of FRPL students. For

Table 7. The Effect of Switching Test Timelines on Schools' Location in the Math and Reading Value-Added Distribution

	Quintile of Percent FRPL				
	(Least)				(Most)
Math Value Added	Q_1	Q_2	Q_3	Q_4	Q_5
Quintile ($\hat{\lambda}^{Spring}$) $>$ Quintile($\hat{\lambda}^{Fall}$)	41.9%	32.9%	26.6%	11.0%	2.6%
Quintile ($\hat{\lambda}^{Spring}$) $=$ Quintile($\hat{\lambda}^{Fall}$)	54.2%	55.5%	59.1%	60.6%	43.5%
Quintile ($\hat{\lambda}^{Spring}$) $<$ Quintile($\hat{\lambda}^{Fall}$)	3.9%	11.6%	14.3%	28.4%	53.9%
Reading Value Added	Q_1	Q_2	Q_3	Q_4	Q_5
Quintile ($\hat{\lambda}^{Spring}$) $>$ Quintile($\hat{\lambda}^{Fall}$)	41.9%	35.5%	27.3%	14.80%	3.3%
Quintile ($\hat{\lambda}^{Spring}$) $=$ Quintile($\hat{\lambda}^{Fall}$)	51.6%	48.4%	51.3%	58.1%	52.6%
Quintile ($\hat{\lambda}^{Spring}$) $<$ Quintile($\hat{\lambda}^{Fall}$)	6.5%	16.1%	21.4%	27.1%	44.2%

example, 42 percent of schools in the bottom quintile of FRPL are in higher quintile of spring-to-spring value-added than fall-to-spring, 54 percent are in the same quintile, and only 4 percent are in a lower quintile.

Limitations

Our paper raises important equity questions related to the design of school accountability policies. Nevertheless, there are a number of constraints that should be kept in mind when interpreting our results. First, NWEA's MAP assessment is not the state-selected accountability test, and is intended to serve as a formative tool for teachers and schools. As a result, students and schools may not take the exam seriously. Over the past decade, however, districts in this southern state have continued to purchase and administer the MAP assessment to their students in grades 2 through 8. It is likely that at least some schools value the information provided by this assessment. Teachers are also less likely to "teach to the test" for a low-stakes assessment, and the MAP assessment may actually provide a better snapshot of students' abilities than a high-stakes exam. Not only does the NWEA data have the appropriate statistical properties for this type of analysis, to our knowledge it is the only longitudinal statewide data that includes fall and spring tests across elementary and middle school grades.

Second, our analysis relies on data from only one state. The demographics in our sample closely resemble many other southern states, although we cannot rule out that our results are driven by students' state-specific idiosyncratic educational experiences. For example, it is also possible our results are driven by the unique interaction between the state's standards and the MAP assessment. The NWEA aligns MAP items to each state's standards, but until the results are replicated in other states, the external generalizability of our analysis may be limited.

Third, our lack of student-level poverty data prohibits us from ruling out student poverty as the source of bias in our analysis, and not students' summer learning. We do know that at the school level, the percentage of minority students is correlated with the percent of FRPL students in a school at $r = 0.65$. Our student-level race indicators, therefore, likely grab at least part of the variation between poverty and achievement. Furthermore, as noted above, the vast majority of students' summer learning is either explained by their prior spring achievement or unobserved factors; variation in socioeconomic status and students' summer experiences explains only 3 to 5 percent of the variation in students' summer learning (Gershenson and Hayes 2016). Furthermore, although not shown in the paper, our results are qualitatively similar if we use quintiles of percent non-white or a principal component of aggregate student demographics throughout our analysis instead of quintiles of aggregate FRPL data.

Lastly, the schools in our analysis did not actually operate under an accountability system that used VAMs to hold them accountable for student outcomes. There are potentially important differences between studying the properties of VAMs when they are not part of an accountability system versus studying their properties when stakes are attached to their outcomes. It is unclear how the schools' knowledge of their performance under an accountability system that used VAMs would alter the results of this paper. It is known, however, that high-stakes tests, relative to low-stakes tests, generally exacerbate achievement differences between minority/white and low-income/wealthy students (c.f. Steele and Aronson's 1995 discussion of stereotype threats), potentially making the results of our analysis a lower bound of the true problem.

6. DISCUSSION AND CONCLUSION

There are a number of important takeaways from our paper. First, although there have been improvements in the design and implementation of school accountability policies, the efficacy of these policies is potentially limited with the continued reliance on a spring-to-spring test timeline. The incorporation of a fall test into the typical accountability system mitigates the potential for the summer period to bias school-level VAMs. In the case for school VAMs, students' fall achievement serves as a summer-free achievement baseline, capturing students' knowledge at the start of the school year. The move to a fall-to-spring test timeline, along with the move to computer adaptive tests, also has the added benefit of providing teachers with information about their students' current achievement levels, and the amount of skills and knowledge lost over the summer.

Second, our results are just another example of many considerations policy makers and researchers must make when using VAMs to hold teachers and schools accountable. On the one hand, the evidence is growing that VAMs may provide unbiased estimates of the effects of teachers and schools on students' achievement (Kane and Staiger 2008; Kane et al. 2013; Chetty, Friedman, and Rockoff 2014a; Deming, 2014). On the other hand, there are important sources of bias that muddy the inferences one can make from these models. Examples of these include the repeated tracking of students to teachers within schools (Horvath 2015), the correlation between student achievement growth and teacher assignment (Rothstein 2015), and now summer learning loss. Furthermore, it is important to keep in mind that observational/behavioral measures of educator quality show similar patterns of bias (Whitehurst, Chingos, and Lindquist 2014). The results of these studies do not necessarily negate the use of VAMs in education research and practice but they do suggest we should carefully consider the tradeoffs associated with their use to allocate rewards and punishments and to make human resource decisions.

Third, the difference in schools' relative value-added ranking between the two test timelines is especially problematic for the new wave of school accountability systems.

Because of the Elementary and Secondary Education Act waivers and now the Every Student Succeeds Act of 2015, many states are incorporating A to F, or similar, grading systems for schools, where the bottom and top 10 to 20 percent will be identified for sanctions and rewards (Polikoff et al. 2014). The movement among quintiles across the two test timelines provides conflicting messages to parents about the quality of their local schools. These conflicting messages have important implications for the public's support for, and satisfaction with, public education, as measured by satisfaction scales, donations to public schools, and the amount of money they are willing to spend on a home (Figlio and Lucas 2004; Figlio and Kenny 2009; Jacobsen, Saultz, and Snyder 2013).[9]

Fourth, even when the summer period is removed from VAMs, there is still a negative correlation between schools' performance and school demographics. It is unclear what the true correlation is between school quality and the schools' political, social, and economic factors, but it is unlikely that it is zero or positive due to labor market preferences, housing preferences, and so on. The results of this paper, and work by Ehlert et al. (2016) and Guarino, Reckase, and Wooldridge (2015), among others, speak to the need for policy makers and educators to first figure out what they want to measure—for example, schools' causal effects on students' achievement or a proportional ranking system—before implementing an accountability policy that uses student achievement growth to hold teachers and schools accountable. Regardless of their choice, it is unlikely that anything short of a fall-to-spring test timeline will remove the bias from students' summer learning.

It is an open question whether schools should be held accountable for all aspects of students' learning or just learning that occurs during the school year. On the one hand, we have known since the Coleman report (see Coleman 1966), and replicated in countless studies, that the vast majority of the variation in students' learning is accounted for by non-school factors, such as family, community, and peer inputs. On the other hand, recent studies have documented long-term impacts of educational inputs (e.g., teachers) on students' non-achievement outcomes, as well as post-secondary attainment and labor market success (Chetty et al. 2011; Jackson 2012; Chetty, Friedman, and Rockoff 2014b). There is a tension between the growing evidence that teachers and schools impact students' lives in ways that holistically change students' behavior (Dobbie and Fryer 2015), and the desire to hold teachers and schools accountable for student outcomes.

The move to a fall-to-spring timeline is also not without its costs. The financial cost of administering standardized tests is minuscule compared with other educational expenditures, averaging well under $100 per student (Chingos 2012). Although the marginal financial cost of administering a second high-stakes exam is small, there are psychological, instructional, and behavioral costs associated with additional tests. A fall-to-spring test timeline may also send the message to schools that they do not need to worry about students' summer activities because they are only held accountable for school-year learning. Additionally, if schools are going to be held accountable for students' learning but not given resources to support their summer activities, the benefits

[9] Although it is less clear that parents' perceptions of school value added are capitalized in house prices (Imberman and Lovenheim 2016).

of providing teachers with a start-of-year achievement snapshot and removing summer learning from accountability models likely outweigh the costs of implementing a second assessment.

The results of our paper, along with Papay (2011) and Gershenson and Hayes (2016), show that policy makers, practitioners, and researchers need to worry about the influence of test timelines on inferences of measures of teacher and school quality commonly used in accountability policies. In our case, we are worried about just one aspect: the influence of students' summer learning and the role it plays on school value added. Whether it is described as bias or not, it is important to understand that students' summer learning does influence the inferences made about school quality in an accountability design that holds schools accountable for student achievement growth, especially schools serving larger shares of students qualifying for the federal FRPL program.

ACKNOWLEDGMENTS

The project was supported by the Kingsbury Data Award funded by the Kingsbury Center at the Northwest Evaluation Association; the Institute of Education Sciences, U.S. Department of Education, through grant R305B100009 to the University of Virginia; and by the Spencer Foundation, through grant 201400162 to North Carolina State University. We are particularly grateful to Dr. Nate Jensen, Research Scientist at the Kingsbury Center, for his help in acquiring the MAP assessment data. We thank the reviewers and editors at *Education Finance and Policy*; participants at the 2013 Association of Education Finance and Policy and 2015 American Educational Research Association meetings; and Seth Gershenson, Eric Parsons, Morgan Polikoff, and Eric Taylor for their input on earlier drafts. The standard caveats apply.

REFERENCES

Alexander, Karl L., Entwisle, Doris R., and Olson, Linda S. 2001. Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis* 23(2):171–191. doi:10.3102/01623737023002171.

Angrist, Joshua, Peter Hull, Parag Pathak, and Christopher Walters. 2017. Leveraging lotteries for school value-added: Testing and estimation. *Quarterly Journal of Economics*. 132(2):871–919.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics*. Princeton, NJ: Princeton University Press.

Atteberry, Allison, and Andrew McEachin. 2016. School's out: Summer learning loss across grade levels and school contexts in the U.S. today. In *The summer slide: What we know and can do about summer learning loss*, edited by K. Alexander, S. Pitcock, and M. Boulay, pp. 35–54. New York: Teachers College Press.

Baker, George. 2000. The use of performance measures in incentive contracting. *American Economic Review* 90(2):415–420. doi:10.1257/aer.90.2.415.

Burkham, David T., Douglas D. Ready, Valerie E. Lee, and Laura F. LoGerfo. 2004. Social-class difference in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education* 77(1):1–31. doi:10.1177/003804070407700101.

Charbonneau, Étienne, and Gregg G. Van Ryzin. 2012. Performance measures and parental satisfaction with New York City schools. *American Review of Public Administration* 41(1):54–65. doi:10.1177/0275074010397333.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane W. Schanzenbach, and Danny Yagan. 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics* 126(4):1593–1660. doi:10.1093/qje/qjr041.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014a. Measuring the impacts of teachers I: Evaluating the bias in teacher value-added estimates. *American Economic Review* 104(9):2593–2632. doi:10.1257/aer.104.9.2593.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014b. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104(9):2633–2679. doi:10.1257/aer.104.9.2633.

Chingos, Matthew M. 2012. *Strength in numbers: State spending on K-12 assessment systems*. Washington, DC: Brown Center on Education Policy at Brookings.

Coleman, James S. 1966. *Equality of educational opportunity*. Available http://files.eric.ed.gov/fulltext/ED012275.pdf. Accessed 7 December 2016.

Cooper, Harris, Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse. 1996. The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research* 66(3):227–268. doi:10.3102/00346543066003227.

Deming, D. J. 2014. Using school choice lotteries to test measures of school effectiveness. NBER Working Paper No. 19803.

Dobbie, Will, and Roland G. Fryer, Jr. 2015. The medium-term impacts of high-achieving charter schools. *Journal of Political Economy* 123(5):985–1037. doi:10.1086/682718.

Downey, Douglas B., Paul von Hippel, and Beckett A. Broh. 2004. Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review* 69(5):613–635. doi:10.1177/000312240406900501.

Downey, Douglas B., Paul von Hippel, and Melanie Hughes. 2008. Are “failing” schools really failing? Using seasonal comparison to evaluate school effectiveness. *Sociology of Education* 81(3):242–270. doi:10.1177/003804070808100302.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael Podgursky. 2013. The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy* 1(1):19–27. doi:10.1080/2330443X.2013.856152.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael Podgursky. 2016. Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy* 30(3):465–500. doi:10.1177/0895904814557593.

Entwisle, Doris R., and Karl L. Alexander. 1992. Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review* 57(1):72–84. doi:10.2307/2096145.

Entwisle, Doris R., and Karl L. Alexander. 1994. Winter setback: School racial composition and learning to read. *American Sociological Review* 59(3):446–460. doi:10.2307/2095943.

Figlio, David N., and Lawrence W. Kenny. 2009. Public sector performance measurement and stakeholder support. *Journal of Public Economics* 93(9–10):1069–1077. doi:10.1016/j.jpubeco.2009.07.003.

Figlio, David N., and Susanna Loeb. 2011. School accountability. In *Handbook in economics: Economics of education*, vol. 3, edited by E. A. Hanushek, S. Machin, and L. Woessmann, pp. 383–421. Amsterdam: Elsevier.

Figlio, David N., and Maurice E. Lucas. 2004. What's in a grade? School report cards and the housing market. *American Economic Review* 94(3):591–604. doi:10.1257/0002828041464489.

Fitzpatrick, Maria D., David Grissmer, and Sarah Hastedt. 2011. What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review* 30(2):269–279. doi:10.1016/j.econedurev.2010.09.004.

Gershenson, Seth. 2013. Do summer time-use gaps vary by socioeconomic status? *American Educational Research Journal* 50(6):1219–1248. doi:10.3102/0002831213502516.

Gershenson, Seth, and Michael S. Hayes. 2016. The implications of summer learning loss for value-added estimates of teacher effectiveness. *Educational Policy*. In press. Advance online publication available doi:10.1177/0895904815625288.

Goldhaber, Dan, and Duncan Dunbar Chaplin. 2015. Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness* 8(1):8–34. doi:10.1080/19345747.2014.978059.

Goldhaber, Dan, and Michael Hansen. 2013. Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica* 80(319):589–612. doi:10.1111/ecca.12002.

Greene, William H. 2003. *Econometric analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall.

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge. 2015. Can value-added measures of teacher performance be trusted? *Education Finance and Policy* 10(1):117–156. doi:10.1162/EDFP_a_00153.

Heyns, Barbara. 1978. *Summer learning and the effects of schooling*. New York: Academic Press.

Holmstrom, Bengt, and Paul Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law Economics and Organization* 7(Special Issue):24–52. doi:10.1093/jleo/7.special_issue.24.

Horvath, Hedvig. 2015. Classroom assignment policies and implications for teacher value-added estimation. Unpublished Paper, University College London.

Imberman, Scott A., and Michael F. Lovenheim. 2016. Does the market value value-added? Evidence from housing prices after a public release of school and teacher value-added. *Journal of Urban Economics* 91:104–121. doi:10.1016/j.jue.2015.06.001.

Jacobsen, Rebecca, Andrew Saultz, and Jeffrey W. Snyder. 2013. When accountability strategies collide: Do policy changes that raise accountability standards also erode public satisfaction? *Educational Policy* 27(2):360–389. doi:10.1177/0895904813475712.

Jackson, C. Kirabo. 2012. Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. NBER Working Paper No. 18624.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have we identified effective teachers? Validating measures of effective teaching using random assignment. MET Project Research Paper. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607.

Ladd, Helen F., and Randall P. Walsh. 2002. Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review* 21(1):1–17. doi:10.1016/S0272-7757(00)00039-X.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. The intertemporal variability of teacher effect estimates. *Education Finance and Policy* 4(4):572–606. doi:10.1162/edfp.2009.4.4.572.

Meng, Xiao-Li, Robert R. Rosenthal, and Donald B. Rubin. 1992. Comparing correlated correlation coefficients. *Psychological Bulletin* 111(1):172–175. doi:10.1037/0033-2909.111.1.172.

Mihaly, Kata, Daniel F. McCaffrey, J. R. Lockwood, and Tim R. Sass. 2010. Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *Stata Journal* 10(1):82–103.

Palardy, Gregory J., and Luyao Peng. 2015. The effects of including summer on value-added assessments of teachers and schools. *Education Policy Analysis Archives* 23(92):1–26.

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

Polikoff, Morgan S., Andrew J. McEachin, Stephanie L. Wrabel, and Matthew Duque. 2014. The waive of the future? School accountability in the waiver era. *Educational Researcher* 43(1):45–54. doi:10.3102/0013189X13517137.

Prendergast, Canice. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37(1):7–63. doi:10.1257/jel.37.1.7.

Quinn, David M. 2015. Black-white summer learning gaps: Interpreting the variability of estimates across representations. *Educational Evaluation and Policy Analysis* 37(1):50–69. doi:10.3102/0162373714534522.

Reardon, Sean F., and Stephen W. Raudenbush. 2009. Assumptions of value-added models for estimating school effects. *Education Finance and Policy* 4(4):492–519. doi:10.1162/edfp.2009.4.4.492.

Rothstein, Jesse. 2015. Revisiting the impacts of teachers. Unpublished paper, University of California, Berkeley.

Rothstein, Richard, Rebecca Jacobsen, and Tamara Wilder. 2008. *Grading education: Getting accountability right*. Washington, D.C., and New York: Economic Policy Institute and Teachers College Press.

Smith, Marshall, and Jennifer A. O'Day. 1991. Systemic school reform. In *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association*, edited by S. H. Fuhrman and B. Malen, pp. 233–267. New York: Falmer Press.

Steele, Claude M., and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology* 69(5):797–811. doi:10.1037/0022-3514.69.5.797.

Todd, Petra, and Kenneth I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal* 113(February):F3–F33. doi:10.1111/1468-0297.00097.

Whitehurst, Grover (Russ) J., Matthew M. Chingos, and Katharine M. Lindquist. 2014. *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at Brookings.

APPENDIX A: ESTIMATION PROCEDURE

In this appendix, we provide more detail on the estimation procedure for our value-added model (VAM). In order to avoid multicollinearity when estimating a DOLS VAM, as in equation 1, the researcher needs to pick a holdout school or reference group. A common practice is to omit an arbitrary school from the model as the holdout school. However, the value-added estimates are then centered on this arbitrary holdout school. Unless the holdout school is of particular interest, the point-estimate of schools' value added is not meaningful. Instead, one can use a sum-to-zero constraint, which centers the estimates around the sample grand mean (see Mihaly et al. 2010 for more detail). The grand-mean centering approach adds qualitative meaning to the value-added estimates, which now represent the difference between a given school's performance and the group average. We used the Stata program FELSDVREGDM to generate our centered value-added estimates (Mihaly et al. 2010) and their standard errors, but one can readily generate these estimates using standard fixed-effect procedures in any statistical software program.

It is a two-step process to estimate a grand-mean centered VAM using any statistical software program. First, run a traditional school fixed-effects model holding out an arbitrary school and store the estimated coefficients on the independent variables (except for the school indicator variables)—that is, the coefficients from equation 1. Second, use the following formula to generate school value-added estimates from the stored coefficients and independent and dependent variables:

$$\hat{\boldsymbol{\lambda}} = \left(\bar{Y}_s - \bar{\boldsymbol{X}}_s\hat{\boldsymbol{\beta}}\right) - \left(\bar{Y}_. - \bar{\boldsymbol{X}}_.\hat{\boldsymbol{\beta}}\right), \tag{A.1}$$

where $\bar{Y}_s$ and $\bar{\boldsymbol{X}}_s$ are school averages of the dependent and independent variables in equation 1, $\bar{Y}_.$ and $\bar{\boldsymbol{X}}_.$ are the unweighted averages of $\bar{Y}_s$ and $\bar{\boldsymbol{X}}_s$, and $\hat{\boldsymbol{\beta}}$ is a vector of all the within-school coefficients from equation 1.[10]

We also see from equation A.1 that the bias in schools' value added enters the estimation equation through $\hat{\boldsymbol{\beta}}$. We now show how the bias formula shown in equation 3 can be derived from equation A.1. To see this more clearly, we will compare the estimates of $\hat{\boldsymbol{\beta}}$ from a short DOLS model ($\hat{\boldsymbol{\beta}}^{Short}$, which does not include students' summer learning) and a long DOLS model ($\hat{\boldsymbol{\beta}}^{Long}$, which does include students' summer learning) (Greene 2003; Angrist and Pischke 2009). We know from the omitted variable bias formula $\boldsymbol{\beta}^{Short} = \boldsymbol{\beta}^{Long} + bias(\hat{\boldsymbol{\lambda}})$, where, ignoring students' summer learning in the off-subject for the moment, $bias(\hat{\boldsymbol{\lambda}}) = \frac{Cov(\boldsymbol{X}^*_{ist}, SL^*_{ist})}{Var(\boldsymbol{X}^*_{ist})}\theta_1^{Summer}$. Substituting this back into equation A.1 and rearranging we obtain the following for short and long value-added specifications:

$$\hat{\boldsymbol{\lambda}}^{Short} = \left(Y_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}}^{Long}\right) - \left(Y_. - \boldsymbol{X}_.\hat{\boldsymbol{\beta}}^{Long}\right) bias\left(\hat{\boldsymbol{\lambda}}\right) \tag{A.2a}$$

$$\boldsymbol{\lambda}^{Long} = \left(Y_s - \boldsymbol{X}_s\hat{\boldsymbol{\beta}}^{Long}\right) - \left(Y_. - \boldsymbol{X}_.\hat{\boldsymbol{\beta}}^{Long}\right) - \left(SL_s - SL_.\right)\theta_1^{Summer}. \tag{A.2b}$$

If we subtract $\hat{\boldsymbol{\lambda}}^{Long}$ from $\hat{\boldsymbol{\lambda}}^{Short}$, similar to equation 3, we get:

[10] For the sake of brevity, we use X to denote any independent variables included in the model, not just student characteristics.

$$\hat{\boldsymbol{\lambda}}^{Short} - \hat{\boldsymbol{\lambda}}^{Long} = \left[\left(\overline{SL}_s - \overline{SL}_. \right) - \left(\bar{\boldsymbol{X}}_s - \bar{\boldsymbol{X}}_. \right) \frac{Cov\left(\boldsymbol{X}^*_{ist}, SL^*_{ist}\right)}{Var\left(\boldsymbol{X}^*_{ist}\right)} \right] \theta_1^{Summer}. \tag{A.3}$$

Rearranging equation A.3, we get

$$\left[\left(\overline{SL}_s - \bar{\boldsymbol{X}}_s \frac{Cov\left(\boldsymbol{X}^*_{ist}, SL^*_{ist}\right)}{Var(\boldsymbol{X}^*_{ist})} \right) - \left(\overline{SL}_. - \bar{\boldsymbol{X}}_. \frac{Cov\left(\boldsymbol{X}^*_{ist}, SL^*_{ist}\right)}{Var(\boldsymbol{X}^*_{ist})} \right) \right] \theta_1^{Summer},$$

which is equivalent to the bias term in equation 3.[11]

APPENDIX B: SCORE PROJECTION

In order to project scores to the first and last day of the school calendar, we combine information from NWEA test results in the fall and the spring, the date of those tests, and knowledge of the southern state's school calendars. For data identification reasons, it was not possible to connect individual school districts to their specific school-year calendars. Nevertheless, beginning in August 2007, this southern state adopted new statewide legislation that specified consistent school start and end dates. We have access to an overview of all school district calendars from 2010–11 through 2013–14, and we can therefore examine the extent to which school districts actually used uniform start and end dates (district level calendars are no longer available prior to 2010–11). In the four years of calendar overviews that we have, it appears that the majority of this southern state's districts use the same school year start and end dates that are described in the legislation: School typically starts on the third Monday of August, and the last day of school falls on the first Thursday of June. Though not every district follows this exact schedule, 55 percent do. In addition, 96 percent of districts' start and end dates fall within three days of these standardized dates. We therefore make a reasonable assumption that districts followed this state-mandated school calendar throughout the panel, and we can infer the school year start date (third Monday of August) and the school year end date (first Thursday in June) for all districts.

Although school year start and end dates are relatively consistent in our sample, the dates on which students took the NWEA tests are not. In the ideal scenario, students would have taken their NWEA tests precisely on the first day of school and the last day of school so that all time between the spring and fall tests was entirely summer. This is obviously not the case. Given that extant research suggests that students learn at a linear rate (Fitzpatrick, Grissmer, and Hastedt 2011), we follow Quinn (2015) to project scores for individual students for what they would have been on the first day of school (e.g., the third Monday in August) and the last day of school (e.g., the first Thursday in June each year).

In order to project the estimated NWEA RIT scores for each student, we calculate the average daily learning rate between each student's fall and spring NWEA test administrations by dividing the change in score by the number of days between the two tests. We then calculate both the number of school days between the start of the school year and each student's fall NWEA test, as well as the number of days of school between each student's spring NWEA and the end of the school year. To project scores to

[11] Note that $\left[\left(\overline{SL}_s - \overline{\boldsymbol{X}}_s \frac{Cov\left(\boldsymbol{X}^*_{ist}, SL^*_{ist}\right)}{Var(\boldsymbol{X}^*_{ist})} \right) - \left(\overline{SL}_. - \overline{\boldsymbol{X}}_. \frac{Cov\left(\boldsymbol{X}^*_{ist}, SL^*_{ist}\right)}{Var(\boldsymbol{X}^*_{ist})} \right) \right] = \hat{\boldsymbol{\lambda}}^{Summer} = \frac{Cov\left(\boldsymbol{\delta}^*_s, SL^*_s\right)}{Var(\boldsymbol{\delta}^*_s)}$.

the start of the school year, we subtract from the student's observed fall score his or her individual daily learning rate multiplied by the number of days between the third Monday of August and the testing date (we follow the analogous procedure for projecting scores to the last day of school).

We find that the observed and projected RIT scores are correlated at 0.9913 (pooled across subjects and fall/spring), and we obtain a root mean squared error of 3.35 when we regress students projected scores on their actual scores. We are therefore confident that the projections do not bias the results presented in this paper.

APPENDIX C: REGRESSIONS

Table C.1. School Value-Added Regressions

	Math		Reading	
	Fall-to-Spring	Spring-to-Spring	Fall-to-Spring	Spring-to-Spring
Prior reading score	0.268***	0.208***	0.319***	0.454***
	(0.002)	(0.001)	(0.004)	(0.001)
Prior math score	0.584***	0.648***	0.412***	0.293***
	(0.005)	(0.001)	(0.003)	(0.001)
Mobile student (within year)	−1.412***	−1.641***	−1.180***	−1.406***
	(0.110)	(0.086)	(0.116)	(0.090)
Mobile student (between years)	−0.461***	−0.322***	−0.519***	−0.352***
	(0.078)	(0.055)	(0.075)	(0.057)
Black student	−2.600***	−2.044***	−1.903***	−1.628***
	(0.060)	(0.033)	(0.055)	(0.035)
Hispanic student	−0.350***	−0.303***	−1.646***	−1.331***
	(0.091)	(0.059)	(0.097)	(0.062)
Other student	0.313**	0.605***	−0.588***	−0.222**
	(0.106)	(0.067)	(0.085)	(0.070)
School enrollment	−0.386	−0.493*	−0.213	−0.309
	(0.664)	(0.239)	(0.514)	(0.250)
% of Mobile students (within)	2.224	4.656**	0.856	2.35
	(3.887)	(1.613)	(3.569)	(1.687)
% of mobile student (between)	0.501	−0.524	−0.872	−1.875***
	(1.158)	(0.454)	(1.191)	(0.475)
% FRPL students	0.718	1.237**	−0.514	−0.122
	(1.077)	(0.403)	(0.934)	(0.422)
% of Black Students	0.011	2.989**	0.700	3.491***
	(2.616)	(0.932)	(2.002)	(0.974)
% of Hispanic Students	−0.447	0.476	4.202	4.736**
	(4.868)	(1.640)	(3.827)	(1.715)
% of other students	2.139	6.027***	0.737	3.229**
	(3.282)	(1.138)	(2.478)	(1.190)
Constant	47.780***	41.488***	66.992***	59.981***
	(5.035)	(1.707)	(3.911)	(1.785)
Adjusted R^2	0.600	0.636	0.487	0.520
Total number of student/years	641,099	641,099	641,099	641,099
Number of schools	774	774	774	774

Note: This model also includes grade fixed-effects, time fixed-effects, and the share of students in a given grade within a school.

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.