The Impact of Summer Learning Loss on Measures of School Performance*

Andrew McEachin[†]

Allison Atteberry[‡]

July 13, 2015

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.

Keywords: Accountability, Value-Added Model, Education Policy, Inequality.

JEL Classification Numbers: H75, I24, I28.

[*] The project was supported in part by the Kingsbury Data Award funded by the Kingsbury Center at the NWEA. We are particularly grateful for Dr. Nate Jensen, Research Scientist at the Kingsbury Center, for his help in acquiring the MAP assessment data. We are also thankful to the helpful comments from participants at the 2013 Association of Education Finance and Policy meetings, as well as Eric Taylor, Seth Gershenson, Eric Parsons, and Morgan Polikoff. The research was also supported by 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. The standard caveats apply.

[†] Corresponding author: RAND Corporation, 1776 Main Street, Santa Monica, CA 90401, mceachin@rand.org

[‡] School of Education, #213, University of Colorado Boulder, 249 UCB, Boulder, CO 80309

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 the schools' purview. Policy-makers at all levels are experimenting with new approaches to this accountability challenge. These advancements, often called value-added models (VAMs), compare the aggregate performance of schools after conditioning on student and school characteristics that are assumed beyond the control of educators and administrators (Castellano & Ho, 2013; Ehlert, Koedel, Parsons,& Podgursky, 2014; Guarino, Reckase, & Wooldridge, 2015; Ladd & Walsh, 2002; Reardon & Raudenbush, 2009; Todd & Wolpin, 2003).

The validity of VAMs to produce measures of school performance rests on a number of assumptions, many of which have been explicated and probed in existing work (Guarino, Reckase, & Wooldridge, 2015; Reardon & Raudenbush, 2009; Todd & Wolpin, 2003). One important, but often ignored, assumption posits that the use of annual test scores, usually administered each spring, measures the amount of learning attributable to a school. However, a student's summer vacation constitutes approximately a quarter of the days in the spring-to-spring testing window. The typical spring-to-spring testing window potentially conflates the amount of learning that occurs during the school-year with the learning that occurs during the summer, largely outside of the schools' control.

The extant research on students' summer experiences suggest 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, while minority and/or disadvantaged children experience losses (Alexander, Entwisle, & Olson, 2001; Atteberry & McEachin, 2015; Downey, Von Hippel, & Broh, 2004; Gershenson & Hayes, 2013)–the negative impact of summer on lower socioeconomic students is often referred to as “summer setback” or “summer learning loss.” The role differential summer setback plays in estimating measures of school quality is further complicated by the systemic sorting of students to schools based on student social and economic characteristics. However, even though a majority of schools are now held accountable for students’ achievement growth from states’ accountability policies under the the federal ESEA waivers (Polikoff, McEachin, Wrabel, & Duque, 2014), no research to date has examined the potential for summer setback to bias VAMs commonly used in research and school accountability policies.

In this paper we use a unique dataset that contains both fall and spring test scores for students in grades two through eight from a Southern state to evaluate the impact of summer setback on VAMs commonly used in state accountability systems to estimate school quality. Specifically, we ask the following questions:

> **Q1:** What is the magnitude and distribution of bias from students’ differential summer learning in spring-to-spring school value-added models?
>
> **Q2:** How does the bias from students’ differential summer learning affect the relative ranking of school quality as measured by schools’ value-add?

The goal of this paper is to combine the research on school accountability and summer setback by evaluating the impact of alternative test timings on school VAMs. We find students’ summer learning biases typical spring-to-spring VAMs used in school accountability policies, 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 data set used to answer our research questions; Section 4 describes our methodological approach; 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 address market failures due to information asymmetry (Baker, 2000; Figlio& Kenny, 2009; Figlio & Lucas, 2004; Holmstrom & Milgrom, 1991; Ladd & Zelli, 2002; Prendergast, 1999). The former assumes the use of performance incentives will better align educators' behaviors with local, state, or federal standards (Prendergast, 1999; Holmstrom & Milgrom, 1991; Smith & O'Day, 1991). The latter problem aims to infuse the educational marketplace with information about the effect of teachers and schools on students' achievement and other outcomes (Charbonneau & Van Ryzin, 2011; Figlio & Loeb, 2011; Rothstein et al., 2008). In both cases, performance is generally defined in terms of students' achievement on standardized tests, which presumes test scores, despite not capturing every skill deemed important by society, are strongly related to students' future success (i.e., Chetty, Friedman, Hilger, Saez, Schanzenbach, & Yagan, 2011; Chetty, Friedman, & Rockoff, 2014b).

The ability for accountability policies to elicit optimal behavior from educators relies on valid measures of teacher and school performance that accurately reflect actors' efforts. If these measures are too noisy, too rigid, or biased by factors outside the actors' control, the incentives to align behaviors with expectation break down and unintended consequences may emerge. Previous research has examined the sensitivity of value-added models to model specification, measurement error, and year-to-year stability (Elhert et al, 2013b; Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Papay, 2011). To date, few studies have evaluated whether the summer period poses a validity threat to the use of value-added and other growth models (Gershenson & Hayes, 2013; Papay, 2011), or inferences about school quality (Downey, von Hippel, & Hughes, 2008).

There is growing evidence that disadvantaged children have fewer learning opportunities during the summer months than their advantaged counterparts. Analyzing a sample of approximately 3,000 fifth and sixth graders in Atlanta, Heyns (1978) found that the gap between disadvantaged and advantaged children's test scores grew during the summer faster than in the school-year. In a later study of students in Baltimore, Entwisle and Alexander (1992, 1994) found that both socioeconomic and race gaps in reading skills grew at faster rates during the summer. More recently, Downey, Von Hippel, & Broh (2004) found that the socioeconomic and racial/ethnic gaps in reading and math skills were not the product of unequal school systems, but in fact widened primarily during the summer.

In a separate paper using the same NWEA data discussed below, Atteberry and McEachin (2015) also examined the overall patterns of student learning growth from second through ninth grade. We found statistically significant and policy relevant variability in students' summer growth rates across all grades, and this differential summer experience contributes meaningfully to the growing disparities in student outcomes that arise during school-age years. In addition, we also found that minority students typically exhibit greater summer learning loss than their white peers, however student demographics alone explain little of the variability across students in summer learning rates.

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 and learn over summer (Cooper et al., 1996; Downey et al., 2004; Heyns, 1978). For example, Gershenson (2013) found that low-income students watch two more hours of TV per day during the summer than students from wealthier backgrounds. However, Gershenson and Hayes (2014) found that even a rich set of student and family observables explain only three to five percent of the variation in students' summer math and reading learning. Given that much of students' summer learning is unobserved to the researcher, value-added models that conflate the summer period with the school year may inappropriately blame schools with disadvantaged students for this summer loss.

To date, little attention has been paid to the intersection between the summer learning loss literature and the use of VAMs to hold teachers and schools accountable for students' achievement growth, partially due to the ubiquitous spring-to-spring test timeline in large-scale education data sets. To date, we know of only three papers that investigate the role that summers play in teacher and school accountability (Gershenson, & Hayes, 2013; Papay, 2011; Downey, von Hippel, and Hughes, 2008). Although Papay (2011) did not directly investigate the role of summer learning loss in teacher value-added models, he found Spearman rank correlations between teachers' spring-spring and fall-spring math and reading value-added of 0.66 and 0.71. Gershenson and Hayes (2013), using data from the Early Childhood Longitudinal Study of 1998-99 (ECLS-k), not only found similar rank correlations but also found that including rich detail about the students' summer activities (including summer school attendance) and parents' backgrounds in spring-to-spring VAMs did not improve the correlation between teachers' fall-to-spring and spring-to-spring math and reading value-add. Lastly, Downey, von Hippel, and Hughes (2008), also using ECLS data, estimated random-effect growth models and found schools serving larger shares of low-income students were more likely to be in the bottom of the performance distribution when school performance measures are based on annual spring-to-spring rather than fall-to-spring test score growth.

The three 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. However, the studies also raise important unanswered questions. First, it is important to note that the Papay (2011) and Gershenson and Hayes (2013) papers examine teacher effects, rather than school effects. Further, these papers do not investigate the relationship between the spring-to-spring and fall-to-spring value-added discordance and student/school demographics. The three studies also have important data limitations. The studies either rely on a few years of data within one urban district (Papay, 2011), or observe only the summer between kindergarten and first grade for one cohort of nationally representative students (Gershenson & Hayes, 2014; Downey, Hippel, & Hughes, 2008).

The results of our paper address the 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 two through eight over a five-year period. Instead of relying on the summer period between kindergarten and first grade, the grade span used in this study is more representative of the grades typically included in high-stakes accountability. Second, we are the first to evaluate the impact of summer setback on school value-added models which are becoming popular in state and federal accountability policies. Lastly, we not only examine whether summer setback leads to potential misclassifications in rank-ordering of schools' math and reading value-add, but also the types of schools that are most affected by this phenomenon.

3 Data

The data for this study are from Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) assessment. MAP is a computer adaptive test that assesses student performance in math, reading and language arts, and science and is administered to students in all 50 states in the U.S and internationally. 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 RIT scale. The vertical scale allows comparisons of student learning across grades and over time, while the equal-interval scale ensures that a unit increase in a student's score represents the same learning gain across the entire distribution.[1] In sum, the MAP assessment has appropriate measurement properties for a study of school VAMs and summer setback.

[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.

The data for our study come from a Southern state that administered the MAP assessment in the fall and spring for all students in grades three through eight for 2007-8 through 2010-11 school-years. During this time period, the Southern state has administered the MAP as a formative assessment to provide teacher and principals information about their students' start- and end-of-year achievement. We discuss the limitations of using low-stakes assessment data in a later section. Our dataset includes 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 that the omission of these variables will bias our results. As noted above, Gershenson and Hayes (2014) 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, compared to 2.5 percent in our data.

<< Insert Table 1 here >>

The school-level data file is provided by the Common Core of Data through NWEA. These data include the typical set of school-level characteristics, including the percent of FRPL students within a school. The student- and school-level descriptives for the 2010-11 school-year are provided in Table 1. Following Quinn (2014), we utilize schools' calendar information to isolate 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

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

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 value-added models to estimate three-year average school effects on students' math and reading achievement. For the spring-to-spring growth models, we use students' prior spring test score as the control for prior achievement. Since our data window starts in the 2007-08 school-year, the first year we can generate schools' value-add using the spring-to-spring test timeline is 2008-09–using the 2007-08 spring score as the lagged achievement variable. 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 three through eight during the 2008-09 to 2010-11 school-years with approximately 45,000 students per grade per year.

4 Methods

The goal of our analyses in this paper is twofold. We first evaluate whether students' summer learning is a potential source of bias in school value-added models. In the second part, we examine how the rank-ordering of schools' math and reading value-add changes when students' summer learning is removed from the value-added model, and how this change may differentially affect schools serving traditionally underserved students.

4.1 Value-Added Model

We start with an education production function (EPF) for students' current achievement proposed in Todd and Wolpin (2003):

$$Y_{igst} = Y_t[\boldsymbol{X_{igst}}, \boldsymbol{Z_{st}}, Y_{igst,t-1}\{\boldsymbol{X_{igs,t-1}}, \boldsymbol{Z_{s,t-1}}, \boldsymbol{\alpha_i}\}, \epsilon_{igst}]. \tag{1}$$

The current achievement of student i (Y_{igst}), in grade g, school s, and time t is modeled as a function of her *current* observable characteristics ($\boldsymbol{X}_{igst}$) and school inputs ($\boldsymbol{Z}_{st}$); her achievement measured in a prior period ($Y_{igs,t-1}$), which is a function of her past observable characteristics ($\boldsymbol{X}_{igs,t-1}$), school inputs ($Z_{s,t-1}$), and fixed academic endowment and family inputs ($\boldsymbol{\alpha}_i$); and her idiosyncratic error term (ϵ_i). The goal is to obtain a unbiased estimate of her school inputs, specifically schools' impact (or value-add) on students' current learning. As will become clear shortly, it is important to think carefully about the empirical and practical implications in the measurement of ($Y_{igs,t-1}$).

Although students' true education production function is unknown to the researcher, extant research suggests that a simple dynamic OLS model (DOLS) which regresses current achievement on prior achievement, school fixed-effects, and vectors of student and school control variables is robust to a number of sources of potential bias (Deming, 2014; Guarino, Reckase, & Wooldridge, 2015). The DOLS specification assumes the 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 decay at a constant geometric rate.

In this paper, we start with the standard DOLS specification:

$$Y_{igst} = \theta_1 Y_{igs,t-1} + \theta_2 \widetilde{Y}_{igs,t-1} + \boldsymbol{\beta X_{igst}} + \boldsymbol{\zeta Z_{gst}} + \boldsymbol{\lambda \delta_s} + \boldsymbol{\alpha_t} + \epsilon_{igst}, \tag{2}$$

where Y_{igst} is modeled as a linear additively separable function of the spring achievement in the prior school-year $t-1$ in both the same and off-subject, Y_{igst} and $\widetilde{Y}_{igst}$; a vector of student demographic characteristics, X_{igst}, including race, a mobility indicator for whether the student made non-structural school move, an indicator for whether students changed schools within the school-year, and an indicator for the 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_{igst}. The key parameters of interest are, $\boldsymbol{\lambda}$, which capture the average conditional achievement of a school's students over the three year panel.[3] As explained in more detail in Appendix A, the $\boldsymbol{\lambda}$ are estimated using a sum-to-zero constraint to ease interpretation, centering $\boldsymbol{\lambda}$ around the state grand mean. Schools with a positive $\boldsymbol{\lambda}$ have value-add above the state average, and vice versa for schools with a negative $\boldsymbol{\lambda}$. We run the model separately for math and reading. The inclusion of student characteristics (including prior achievement) and school fixed-effects account for the sorting of student to schools by these observable characteristics, getting school effects on students' math and reading achievement net of these background characteristics.

Although questions have been raised about the ability of Model 2 to produce unbiased effects of schools on students' achievement (Rothstein, 2011), it has been shown to replicate school effects in experimental and quasi-experimental conditions (Chetty et al., 2014a;

[3] 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 important differences in schools' value-add by changing the test timeline but it also allows us to see if year-to-year variation in value-add 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-add generates results similar to those presented in this paper when comparing spring-to-spring to fall-to-spring value-add. We find that the year-to-year fluctuation in spring-to-spring value-add 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 (2) that uses multiple years of data to generate a single value-added estimate.

Guarino, Reckase, & Wooldridge, 2015; Kane et al, 2013; Deming, 2014).[4] However, it is still possible that students' summer learning loss is a source of bias in these randomized and quasi-randomized studies of VAMs. For example, Kane and Staiger (2008) and Kane et al. (2013) use within-school random assignment of teachers to students to estimate the bias in non-experimental measures of teacher effectiveness. The within-school comparison of teacher quality removes any between school sources of bias. If students are sorted across schools, but bias comparisons in these studies are restricted to within schools, then these experiments will not pick up the between school bias in teachers' value-added due to students' summer learning. Furthermore, Chetty et al. (2014a) evaluate whether teacher VAMs are biased from the omission of parental income data. However, as noted above, even a rich set of family inputs only explains a small portion (approximately 5 percent) of students' summer learning (Gershenson & Hayes, 2014). It is likely that the teacher quality estimates with and without parental income data are equally biased by students' summer learning. Finally, it is possible to find no bias on average, but for the bias to affect a non-trivial share of teachers or schools (cf. Goldhaber, Cowan, & Welch, 2013).

Given that most state-wide testing systems use a spring-to-spring test timeline, researcher and policy-makers have to use students' prior spring achievement ($Y_{igs,t-1}^{Spring}$) to capture the effect of prior inputs on current achievement. Roughly three months of the time between $Y_{igs,t-1}^{Spring}$ and Y_{igst}^{Spring} occurs outside of school, potentially conflating school-year and "summer learning loss" in the estimates of $\boldsymbol{\lambda}$. If instead we wanted to measure the impact of schools' on school-year learning, we would want to condition on students' achievement

[4] Another approach is to estimate school value-added in two steps. The first step estimates model (2) without school fixed-effects and uses the estimate regression coefficients to generate estimate student-level residuals. The second step is to create school-level averages of the residuals as estimate schools' value-add. This model, often called the Average Residual (AR) approach (see Elhert et al. (2014) or Guarino, Reckase, and Wooldridge (2015) for more detail) has the benefit of mechanically removing any correlation between the variables included in the model and schools' value-added. Koedel and Li (Forthcoming) argue that this independence between school demographics and value-added, or what they call proportionality, may enhance the efficiency of education production process. However, since the model does not partial out student and school covariates while simultaneously estimating school effects, the AR approach does not produce unbiased estimates under non-random student assignment (Guarino, Reckase, & Wooldridge, 2015). In order to simplify our exposition, we do not present the results from the AR. However, the results are qualitatively similar when using the DOLS or AR to generate school effects and are available upon request.

on the first day of school Y_{igst}^{Fall} instead of $Y_{igs,t-1}^{Spring}$. For example, we define students' fall achievement as $Y_{igst}^{Fall} = SL_{igst} + Y_{igs,t-1}^{Spring}$, where SL_{igst} captures students' summer learning. Only when $SL_{igst} = 0$ will students' fall achievement equal their prior spring achievement. If $SL_{igst} \neq 0$ students' summer learning is part of the error term in Model 2:

$$Y_{igst}^{Spring} = \theta_1 Y_{igs,t-1}^{Spring} + \theta_2 \widetilde{Y}_{igs,t-1}^{Spring} + \boldsymbol{\beta X_{igst}} + \boldsymbol{\zeta Z_{gst}} + \boldsymbol{\lambda \delta_s} + \boldsymbol{\alpha_t} + (\theta_1 SL_{igst} + \theta_2 \widetilde{SL}_{igst} + \eta_{igst}). \quad (3)$$

If students' summer learning (SL_{igst} and $\widetilde{SL}_{igst}$) is correlated with the independent variables in Model 3, 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.[5]

Whether explicit, one of two assumptions are made when using a spring-to-spring test timeline to estimate Model 2:

> **Assumption 1:** Students' summer learning is independent of school quality, $Cov(\boldsymbol{\delta_s}, SL_{igst}) = Cov(\boldsymbol{\delta_s}, \widetilde{SL}_{igst}) = 0$.
>
> **Assumption 2:** School quality, and student and school characteristics are the main determinants of students' summer learning, $SL_{igst} = \theta_1 Y_{igs,t-1}^{Spring} + \theta_2 \widetilde{Y}_{igs,t-1}^{Spring} + \boldsymbol{\beta X_{igst}} + \boldsymbol{\zeta Z_{gst}} + \boldsymbol{\lambda \delta_s} + \boldsymbol{\alpha_t} + \xi_{igst}$, where $Cov(\boldsymbol{\delta}_s, \xi_{igst}) = 0$.

The first assumption simply states that students' summer learning is independent of school quality, and therefore its presence in the error term of (3) *does not* bias Model (2). As discussed in more detail in Appendix A, this also means that all of the covariates included in Model (2) are independent of students' summer learning. Extant research suggests that this is likely an untenable assumption (cf. Atteberry, & McEachin, 2015). The second assumption states that students' summer learning is a direct function of school quality, and therefore it should not be included in Model (2) since students' summer learning is

[5] We ignore other sources of noise for the moment (e.g., test measurement error, sampling variation).

included in the achievement growth between time $t-1$ and t. Although it is quite possible that schools' have a direct impact on students' summer learning, especially since teacher and school quality predict long-run differences in students' outcomes, it is very unlikely that, with the exception of idiosyncratic noise in the year-to-year variation in students' summer learning, school quality is the determining factor in students' summer learning.

Extant research, and the results in this paper, show that students' summer learning does vary across schools in systematic ways. In this paper, we assume that the majority of this variation is outside of the schools' control. We also assume for the moment students' summer learning is the only source of bias in Model (3)[6] and estimate the bias in schools' spring-to-spring math and reading value-add from students' summer learning in two ways.

First, we estimate whether students' summer learning biases schools' math and reading value-add by relying on students' prior spring (instead of current fall) achievement as their lagged achievement in Model 2. We generate spring-to-spring and fall-to-spring estimates of schools' math and reading value-add from Model 2, using the same vector of student and school covariates in each model. If, on average, schools' spring-to-spring value-add is an unbiased estimate of schools fall-to-spring value-add, the coefficient of a simple OLS regression of schools' fall-to-spring value-add on their schools' spring-to-spring value-add should produce a coefficient statistically indistinguishable from 1 (Chetty et al, 2014a; Deming, 2014; Kane & Staiger, 2008). We test this by estimating a simple OLS regression of schools' fall-to-spring VA on spring-to-spring VA, and testing whether the coefficient on the spring-to-spring VA equals one (i.e., $\hat{\phi} = 1$).

[6] Our assumption that (e.g., $E[\eta_{igst}|Y^{Spring}_{igs,t-1}, \widetilde{Y}^{Spring}_{igs,t-1}, \boldsymbol{X_{igst}}, \boldsymbol{Z_{st}}, \boldsymbol{\delta_s}, SL_{igst}, \widetilde{SL}_{igst}] = 0$) follows the recent strong claims about the ability of a simple value-added model to estimate the causal impact of teachers and schools on students' achievement (e.g., Chetty et al, 2014a; Deming, 2014). However, it is possible that other sources of bias exist in schools' value-add other than students' summer learning loss. We argue that while possible, it's likely that this source of bias is similar in both a spring-to-spring and fall-to-spring test timeline. For example, if students sort to schools based on factor unobserved to the researcher, this bias would likely equally affect value-added from either timeline. If this is true, our results can be interpreted as the relative bias caused by summer learning loss between two models.

Second, we estimate the unit specific bias in schools' math and reading value-add as

$$Bias(\hat{\boldsymbol{\lambda}}) = E[\hat{\boldsymbol{\lambda}}] - \boldsymbol{\lambda} = \frac{Cov(\delta_s^*, SL_{igst}^*)}{Var(\delta_s^*)}\theta_1^{Summer} + \frac{Cov(\delta_s^*, \widetilde{SL}_{igst}^*)}{Var(\delta_s^*)}\widetilde{\theta}_2^{Summer}, \tag{4}$$

where δ_s^* and SL_{igst}^* and $\widetilde{SL}_{igst}^*$ are residualized school indicators and measures of students' summer learning, respectively, using the student and school covariates from Model 2. The terms θ_1^{Summer} and $\widetilde{\theta}_2^{Summer}$ are coefficients for SL_{igst} and $\widetilde{SL}_{igst}$ if they were included as variables in Model 3. Atteberry and McEachin (2015) found these terms to be positive.

The terms $\frac{Cov(\delta_s^*, SL_{igst}^*)}{Var(\delta_s^*)} = \boldsymbol{\lambda}^{Summer}$ and $\frac{Cov(\delta_s^*, \widetilde{SL}_{igst}^*)}{Var(\delta_s^*)} = \widetilde{\boldsymbol{\lambda}}^{Summer}$ in (4) capture the average amount of summer learning within school s conditional on students' prior achievement and student and school covariates. We estimate $\boldsymbol{\lambda}^{Summer}$ and $\widetilde{\boldsymbol{\lambda}}^{Summer}$ from separate auxiliary DOLS regressions of Model 2 with students' summer learning as the dependent variable[7]:

$$SL_{igst} = \theta_1 Y_{igs,t-1}^{Spring} + \theta_2 \widetilde{Y}_{igs,t-1}^{Spring} + \boldsymbol{\beta X_{igst}} + \boldsymbol{\zeta Z_{st}} + \boldsymbol{\lambda}^{Summer}\boldsymbol{\delta_s} + \boldsymbol{\alpha_t} + \epsilon_{igst} \tag{5a}$$

$$\widetilde{SL}_{igst} = \theta_1 Y_{igs,t-1}^{Spring} + \theta_2 \widetilde{Y}_{igs,t-1}^{Spring} + \boldsymbol{\beta X_{igst}} + \boldsymbol{\zeta Z_{st}} + \widetilde{\boldsymbol{\lambda}}^{Summer}\boldsymbol{\delta_s} + \boldsymbol{\alpha_t} + \epsilon_{igst}. \tag{5b}$$

Similar to how we estimate the school value-added in Model 2, we estimate Models 5a and 5b using a sum-to-zero constraint: Each coefficient represents the amount of conditional summer learning within a school centered around the state grand mean. We conduct separate joint F-tests to evaluate whether $\hat{\boldsymbol{\lambda}}^{Summer} = 0$ and $\hat{\widetilde{\boldsymbol{\lambda}}}^{Summer} = 0$.[8]

We evaluate the conditions under which the bias in (4) is present in (2) in the next section. The bias in (4) is particularly policy relevant if the correlation between the estimated bias and the share of traditionally under-served students within a school is

[7] In reality, we only need to estimate 5a and 5b for one subject. The difference between 5a and 5b for the math and reading value-added investigations is the labeling of *same*- and *off*-subject; the covariates and school indicators remain unchanged.

[8] If one is only interested in the relative ranks of schools' value-add, not the point estimates themselves, then it is only important to test whether $\hat{\lambda}_1^{Summer} = \hat{\lambda}_2^{Summer} = ... = \hat{\lambda}_S^{Summer}$ and $\hat{\widetilde{\lambda}}_1^{Summer} = \hat{\widetilde{\lambda}}_2^{Summer} = ... = \hat{\widetilde{\lambda}}_S^{Summer}$.

negative, penalizing schools for educating students from disadvantaged backgrounds. In the next section, we start our investigation into the potential problems summer setback causes for school accountability policies by estimating both the average and school-specific bias in schools' spring-to-spring math and reading value-add.

5 Results

5.1 Q1: What is the magnitude and distribution of bias from students' differential summer learning in spring-to-spring school value-added models?

In evaluating the bias in schools' spring-to-spring math and reading value-add, we start with the assumption that the fall-to-spring DOLS specification generates unbiased estimates of schools' effects on students' math and reading achievement, and captures the true effect of schools on students' achievement. The question is whether estimates of schools' spring-to-spring value-add, $\hat{\boldsymbol{\lambda}}^{Spring}$, are unbiased estimates of their fall-to-spring value-add, $\hat{\boldsymbol{\lambda}}^{Fall}$. We start our analysis with visual inspection of the relationship between these two measures of schools' math and reading value-add in Figures 2a and 2b. If schools' spring-to-spring math and reading value-add is an unbiased 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. If instead students' summer learning biases $\hat{\boldsymbol{\lambda}}^{Spring}$, then the points will form a circle around the 45° line. Even though the points in Figures 2a and 2b are roughly scattered around the 45° line, the relationship between $\hat{\boldsymbol{\lambda}}^{Fall}$ and $\hat{\boldsymbol{\lambda}}^{Spring}$ does suggest a small systematic bias. The confidence intervals of linear and local polynomial regressions through the scatter plots in 2a and 2b show that, on average, the points do not fall on the 45° line–although the average deviation appears modest.

<< Insert Figures 2a and 2b here >>

We further explore the average relationship between $\hat{\boldsymbol{\lambda}}^{Fall}$ and $\hat{\boldsymbol{\lambda}}^{Spring}$ with simple OLS regressions of $\hat{\boldsymbol{\lambda}}^{Fall} = \phi_0 + \phi_1 \hat{\boldsymbol{\lambda}}^{Spring} + \epsilon_s$, and test whether $\phi_1 = 1$ (separately for math and reading); we also run quantile regressions at the 10th, 25th, 50th, 75th, and 90th quantiles of $\hat{\boldsymbol{\lambda}}_s^{Fall}$. The results of these OLS and quantile regressions are presented in Table 2. For both math and reading, we reject the hypothesis that $\hat{\phi} = 1$ in the OLS specification, and across all various quantiles of $\hat{\boldsymbol{\lambda}}^{Fall}$. The projection bias $(1 - \hat{\phi})$ from using a spring-to-spring test timeline to estimate (2) ranges between 8 and 24 percent depending on the subject and placement of the school in the $\hat{\boldsymbol{\lambda}}^{Fall}$ 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 1 unit increase in math $\sigma_{\boldsymbol{\lambda}^{Spring}}$ (2.44 RIT points). As a result of that policy change, we would expect students within-year learning to increase only $.85 * \sigma_{\boldsymbol{\lambda}^{Spring}}$. 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-add, but do not describe the distribution of the bias.

<< Insert Table 2 here >>

Next, we estimate the components of (4) to generate the school-specific bias in schools' $\hat{\boldsymbol{\lambda}}^{Spring}$ value-add from students' differential summer learning. Recall that $\hat{\boldsymbol{\lambda}}^{Summer}$ are coefficients on school indicator variables estimated from auxiliary regressions (5a) and (5b) of students' summer learning SL_{igst} regressed on the independent variables from (2). These coefficients $(\hat{\boldsymbol{\lambda}}^{Summer})$ capture the average amount of summer learning loss within a school after partialing out students' prior spring achievement and student and school characteristics. We run an F-test on the joint-hypothesis that students' summer learning does not vary across schools, conditional on prior achievement and student and school observable characteristics, in (5a) and (5b), respectively; a failure to reject this hypothesis indicates that there is significant variation in summer learning loss across schools, even conditional on student and school covariates.[9]

<< Insert Table 3 here >>

As shown in the first row of Panel A in Table 3, we reject the null hypothesis that $\hat{\boldsymbol{\lambda}}^{Summer} = 0$ for both math and reading. Furthermore, we construct 95 percent confidence intervals for $\hat{\boldsymbol{\lambda}}^{Summer}$ to evaluate how many schools have negative, zero, and positive estimates of conditional math and reading summer learning loss.[10] Panel A indicates 21 and 26 percent of the schools have negative estimates of aggregate conditional math and reading summer learning and 28 and 32 percent 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-add is met, approximately 50 percent of the schools in our sample would have a different math or reading value-add in a fall-to-spring timeline than a spring-to-spring timeline due to students' differential summer learning.

In order for students' summer learning to bias schools' spring-to-spring value-add, summer learning must not only be non-randomly distributed among schools but it also must have a statistically significant relationship with students' spring achievement, conditional on the covariates in (2). As shown in Panel B, students' same-subject and off-subject summer learning has a positive, statistically significant relationship with students spring achievement. These coefficients are estimated by including students' same- and off-subject summer learning in (3), instead of leaving SL_{igst} and $\widetilde{SL}_{igst}$ in the error term. The results of Panel A and Panel B indicate that even conditional on students' prior spring achievement and student and school covariates, students' summer learning meets the necessary criteria to impose a source of bias in schools' spring-to-spring math and reading value-add. 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'

[9] specifically, we test ($H_o : \lambda_1^{Summer} = \lambda_2^{Summer} = ... = \lambda_s^{Summer} = 0$), separately for math and reading.

[10] Although we report the results for math and reading separately in Panel A, in practice the bias calculation for a given subject (i.e., math) includes the sum of math and reading λ^{Summer}.

spring-to-spring value-add, and that this bias is strongly negatively correlated with the percent of FRPL students ($r = -.61$) and minority students ($r = -.41$) in a school.

<< Insert Figures 3a and 3b here >>

We present graphical representations of the bias in Figures 3a and 3b. Specifically, we plot the kernel density of the bias in schools' math and reading spring-to-spring value-add 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-add 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 the bias in spring-to-spring value-add from students' summer learning loss.

5.2 Q2: How does the bias from students' differential summer learning affect the relative ranking of school quality as measured by schools' value-add?

To answer the second research question we start with a simple examination of the Spearman rank-order correlation among the schools' spring-to-spring and fall-to-spring value-add and aggregate student demographics in Table 4. Here we examine whether the absence of information about each student's starting point in the fall has a large impact on how schools are ranked. The correlation of schools' math and reading rankings between $\hat{\boldsymbol{\lambda}}^{Spring}$ and $\hat{\boldsymbol{\lambda}}^{Fall}$ timeline is relatively stable ($r = .85$). Similar to extant research in teacher value-added (Papay, 2011), the cross-subject rankings, however, are lower (ranging from $r = .4$ to .6). Although our school-level within-subject correlations of $\hat{\boldsymbol{\lambda}}^{Spring}$ and $\hat{\boldsymbol{\lambda}}^{Fall}$ are stronger than the year-to-year correlations of teacher value-added (McCaffrey, Sass, Lockwood, & Mihaly, 2009), it is still possible for the discordance between test-timelines to have important implications for school accountability systems.

<< Insert Table 4 here >>

In the last two rows of Table 4, we show the correlations among schools' math and reading value-add and aggregate student demographics. For both math and reading, the spring-to-spring value-add have strong negative correlations with the percent of FRPL students in a school ($r_{math} = -.51$) and ($r_{reading} = -.60$) and the percent of minority students in a school ($r_{math} = -.50$) and ($r_{reading} = -.56$). Schools' math and reading fall-to-spring value-add, measures that do not include students' summer learning, have weaker negative correlations with the percent of FRPL students ($r_{math} = -.22$) and ($r_{reading} = -.34$) and the percent of minority students in a school ($r_{math} = -.22$) and ($r_{reading} = -.29$). These differences are statistically significant ($p \leq .01$) (Meng, Rosenthal, & 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. However, it is important to understand that altering the test timeline does not change the theoretical education production function for schools; instead, it can be argued that the fall-to-spring test timeline more closely mirrors how researchers, policy-makers, and practitioners view schools' impact on student achievement. We further explore the potential for students' summer learning loss to affect schools' relative standing by examining transition matrices between schools' math and reading $\hat{\boldsymbol{\lambda}}^{Spring}$ and $\hat{\boldsymbol{\lambda}}^{Fall}$ in Tables 5 and 6.

<< Insert Tables 5 and 6 here >>

In Panel A of Tables 5 and 6 we present the transition matrix across quintiles of schools' math and reading spring-to-spring and fall-to-spring value-add. Similar to Gershenson and Hayes (2013), and to a lesser degree Papay (2011), we find non-trivial 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. For example, of the 155 schools in the bottom quintile in $\hat{\boldsymbol{\lambda}}^{Spring}$, 48 of them (31 percent) are in higher quintiles of $\hat{\boldsymbol{\lambda}}^{Fall}$, including 24 in at least the third quintile. Similar patterns emerge when comparing the share of schools in the bottom quintile of $\hat{\boldsymbol{\lambda}}^{Fall}$. Furthermore, there is similar movement for schools in the fourth quintile of either $\hat{\boldsymbol{\lambda}}^{Spring}$ or $\hat{\boldsymbol{\lambda}}^{Fall}$. 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. The 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 matrix 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-add in both timelines, and schools with larger shares of FRPL students are clustered in the bottom quintiles of value-add in both timelines. However, as suggested the negative relationship between bias from students' summer learning loss and school poverty, the relationship between school poverty and school performance is stronger in the spring-to-spring time than the fall-to-spring timeline. For example, schools in Panel B are 2.2 times *more* likely to be in the bottom two quintiles of $\hat{\boldsymbol{\lambda}}^{Fall}$ compared to $\hat{\boldsymbol{\lambda}}^{Spring}$, while schools in Panel B are 1.4 times more likely to be in the bottom two quintiles of $\hat{\boldsymbol{\lambda}}^{Spring}$ compared to $\hat{\boldsymbol{\lambda}}^{Fall}$. 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-add) was independent, each row and column total would sum to approximately 31. The row totals for $\hat{\boldsymbol{\lambda}}^{Spring}$ 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 $\hat{\boldsymbol{\lambda}}^{Fall}$ ranging from 19 to 49 for Panel B and 24 to 41 in Panel C.

We take a final look at how the bias in schools' spring-to-spring value-add from students' summer learning affects the inferences made on school quality in Table 7. Specifically, we look at the share of schools in a higher, equal, or lower quintile of spring-to-spring value-add compared to 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 example, 42 percent of schools in the bottom quintile of FRPL are in higher quintile of spring-to-spring value-add than fall-to-spring, while 54 percent are in the same quintile, and only 4 percent are in a lower quintile.

5.3 Limitations

The results of this paper raise important equity questions related to the design of school accountability policies. However, 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 as seriously. Yet, districts in this Southern state have continued to purchase and administer the MAP assessment to its students in grades two through eight over the past decade. 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 state-wide 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; however, 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 in a school 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 SES and student summer explains only 3 to 5 percent of the variation in students' summer learning (Gershenson and Hayes, 2014). 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 value-added models to hold them accountable for student outcomes. There are potentially important differences between studying the properties of VAMs when educators are not actually held accountable to them 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. Nonetheless, it is unlikely that schools operating under an accountability system that used growth-based VAMs would be able to overcome the bias due to summer periods. In fact, high-stakes tests, relative to low-stake tests, generally exacerbate achievement differences between minority/white and low-income/wealthy students (cf Steele and Aronson's (1995) discussion of stereotype threats), potentially making the results of our analysis a lower bound on the true problem.

6 Discussion and Conclusion

The goal of this paper was threefold. First, we examined whether students summer learning loss (SLL) met the two conditions sufficient to introduce bias into a spring-to-spring school value-added model: 1) SLL is predictive of students' current spring achievement conditional on their past spring achievement, student and school demographics, and school fixed-effects; and 2) SLL is inequitably distributed across schools, conditional on the same variables as 1). We found that both of these conditions are in fact met. Second, we demonstrated that the standard deviation of the bias is roughly 25 percent schools' value-add estimated using a spring-to-spring test timeline. Third, we showed that a spring-to-spring test timeline is more likely to over-identify schools serving larger shares of FRPL students as low performing, and under-identify these schools as high performing. Our paper is the first to evaluate the potential bias from students' differential summer learning introduced into school value-added models using a spring-to-spring test timeline, and the results raise a number of important design and equity questions related to school accountability policies.

The first takeaway is that while there have been improvements in the design and implementation of growth models in school accountability, 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 value-added models. In the case for school value-added models, 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 value-added models to hold teachers and schools accountable. On the one hand, the evidence is growing that value-added models provide unbiased estimates of the effects of teachers and schools on students achievement (Chetty, et al, 2014a; Deming, 2014; Kane et al, 2013; Kane & Staiger, 2008). On the other hand, there are important sources of bias that muddy the inferences of teacher and school quality 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, 2014), and now summer learning loss. Furthermore, it is important to keep in mind that observational measures of educator quality show similar patterns of bias (Whitehurst, Chingos, & Lindquist, 2014). The results of these studies do not necessarily negate the use of value-added models 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. Due to the ESEA waivers, 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). Furthermore, the movement among quintiles across the two test timelines tells parents conflicting messages about the quality of their local school. These conflicting messages have important implications for the public's support for, and satisfaction with, public education. For example, extant research suggests that parents are sensitive to the movement of schools across letter grades, as evidenced in donations to public schools and measures of their satisfaction (Figlio, & Kenny, 2009; Jacobsen, Saultz, & Snyder, 2013).

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 Elhert et al. (2014) and Guarino et al. (2014), among others, speak to the need for policy-makers and educators to first figure out what they want to measure–e.g., school's causal effect 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.

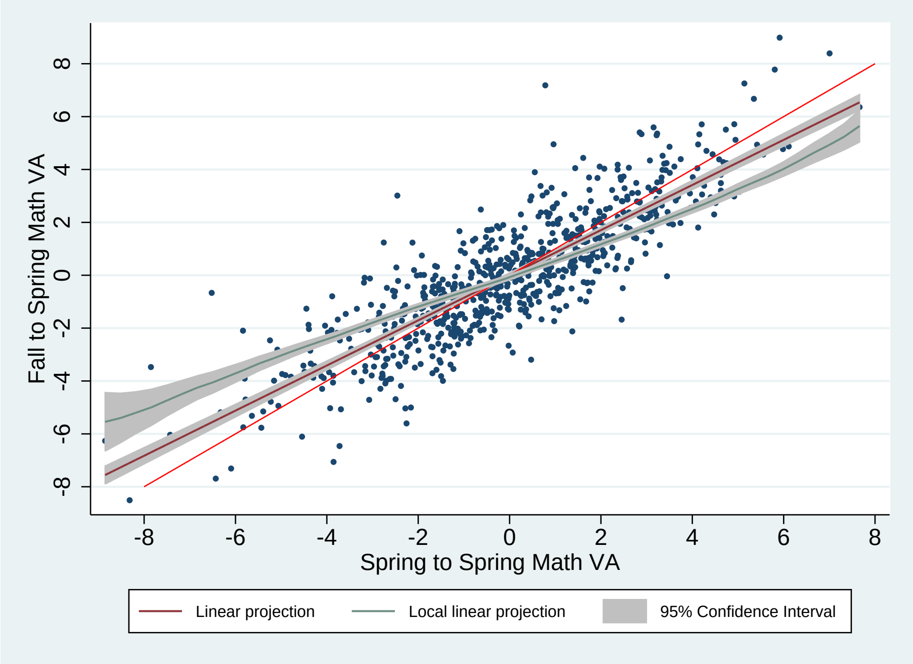
It is an open question whether schools should be held accountable for all aspects of students' learning, or just learning that occurs during the school year. On the one hand we have known since the Coleman report, and replicated in countless studies, that the vast majority of the variation in students' 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; Chetty, Friedman, & Rockoff, 2014b; Jackson, 2012). There is a tension between the growing evidence that teachers and schools impact students' lives in ways that holistically change students' behavior (Dobbie & Fryer, Forthcoming), and the desire to hold teachers and schools accountable for students' 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 to other educational expenditures, averaging well under $100 per student (Chingos, 2012). While the marginal financial cost of administering a second high-stakes exam is small, there are psychological, instructional, and behavioral costs associated with implementing a fall test. The implementation of a fall-to-spring test timeline may also send the message to schools that they do not need to worry about students' summer activities since they are only held accountable for school-year learning. One way around this would be to build incentives around students' fall scores. However, if schools are going to be held accountable for students' learning but not given resources to support their summer activities, the benefit 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.

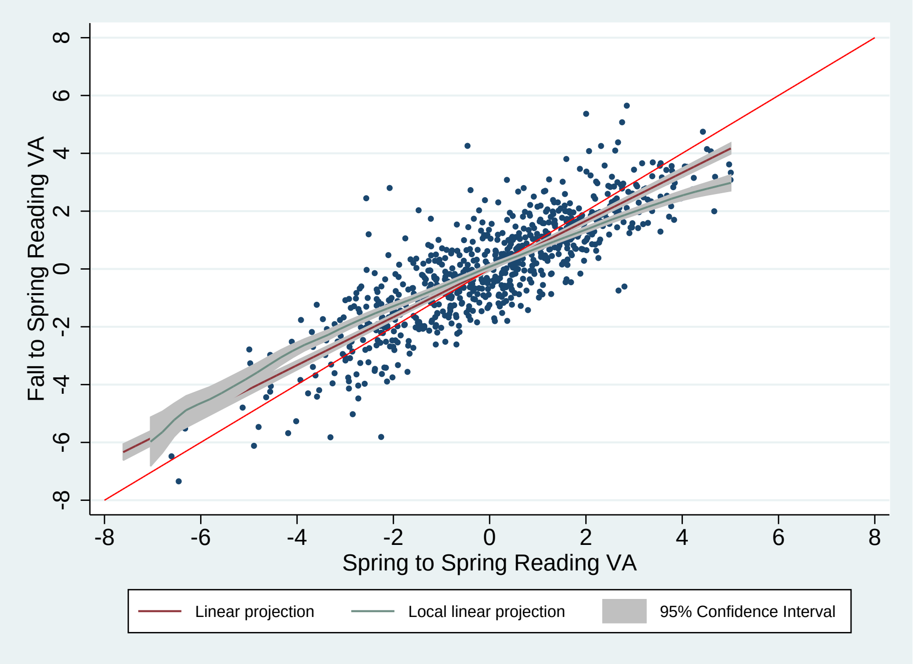
The results of our paper, along with Papay (2011) and Gershsenson and Hayes (2013), show that it is possible to make different inferences about educator and school quality depending on the design of an accountability system. In our case, we are worried about just one aspect: the influence of students' summer learning and the role it plays on school value-added. Whether it is described as bias or not, it is important to understand students' summer learning does influence the inferences made about school quality in an accountability design that holds schools accountable for student achievement growth, especially at the tails of the joint distribution of school demographics and quality.

Figures and Tables

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

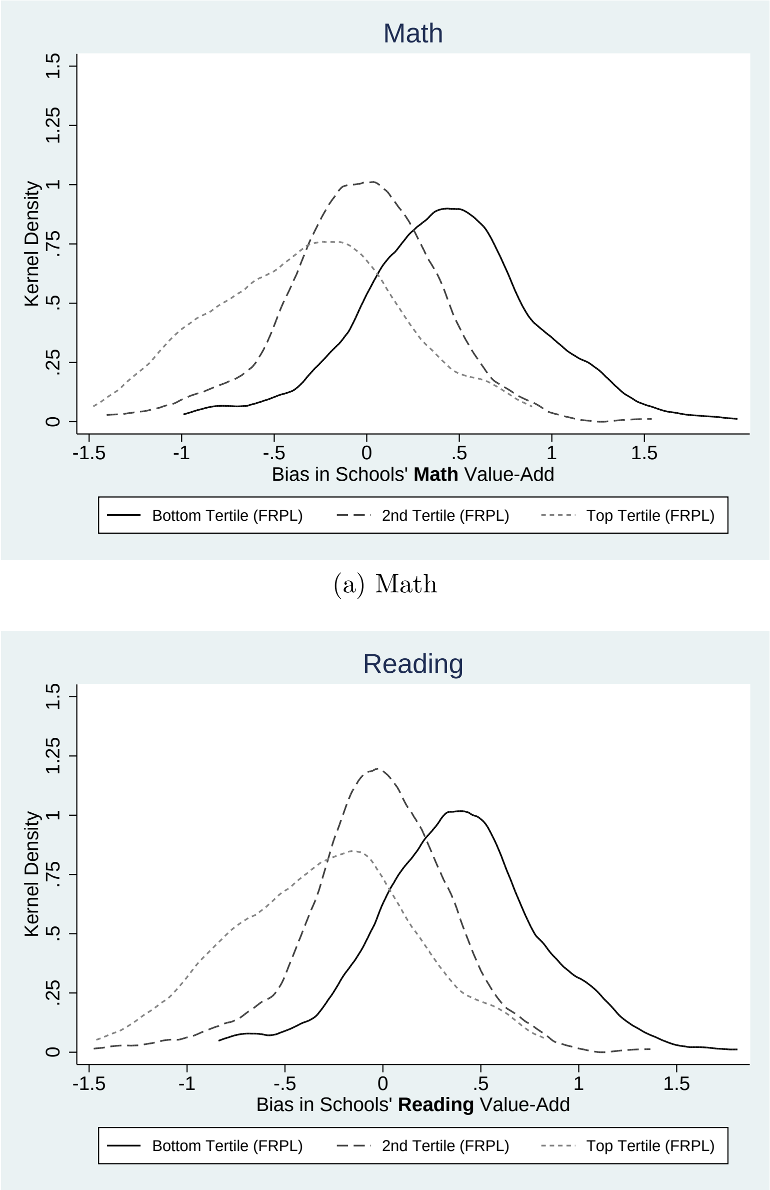


(a) Math



(b) Reading

Figure 2: Kernel Density of the Bias in Schools' Math and Reading Value-Add



(a) Math

(b) Reading

Table 1: Student Demographics for the 2010-11 School Year

Student Demographics	Mean	SD	N
Spring Math MAP Achievement	224.22	19.67	222765
Spring Reading MAP Achievement	214.89	17.04	219665
Fall Math MAP Achievement	212.08	20.16	222765
Fall Reading MAP Achievement	206.07	18.75	222765
Summer Math MAP Loss	-4.21	9.50	163495
Summer Reading MAP Loss	-2.50	11.25	161825
Lagged Spring Math MAP Achievement	216.33	20.48	222765
Lagged Spring Reading MAP Achievement	208.86	18.31	222765
White Student	0.530		222765
Black Student	0.363		222765
Hispanic Student	0.059		222765
Mobile Student (Between School Years)	0.079		222765
Mobile Student (Within School Years)	0.029		222765
3rd Grade	0.171		222765
4th Grade	0.182		222765
5th Grade	0.183		222765
6th Grade	0.160		222765
7th Grade	0.157		222765
8th Grade	0.148		222765
% Hispanic in School	0.063		222765
% Black in School	0.357		222765
% White in School	0.527		222765
% FRPL in School	0.577		222765
Urban School	0.169		222765
Suburban School	0.265		222765
Town School	0.145		222765
Rural School	0.421		222765
School Enrollment	673.17	254.51	222765
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

Table 2: Regression of Schools' Fall-to-Spring VA on Spring-to-Spring VA

			Math			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	OLS	$\tau = .10$	$\tau = .25$	$\tau = .50$	$\tau = .75$	$\tau = .90$
Schools' Spring-to-Spring Math Value-Add	.854***	.919***	.871***	.822***	.823***	.875***
	(.021)	(.031)	(.020)	(.021)	(.029)	(.039)
Adjusted R^2	.725					
# of Schools	774	774	774	774	774	774
P-value ($\phi = 1$)	.000	.010	.000	.000	.000	.001
			Reading			
Schools' Spring-to-Spring Reading Value-Add	.833***	.905***	.848***	.820***	.780***	.767***
	(.022)	(.028)	(.024)	(.020)	(.024)	(.033)
Adjusted R^2	.727					
# of Schools	774	774	774	774	774	774
P-value ($\phi = 1$)	.000	.001	.000	.000	.000	.000

$*p < 0.05, **p < 0.01, ***p < 0.001$

Table 3: Analysis of Potential Bias in Schools' Math and Reading Value-Add from Students' Summer Learning

	Math	**Reading**
Panel A: $\frac{Cov(\delta_s^*, SL_{igst})}{Var(\delta_s^*)}$		
$P - value : \lambda_1^{Summer} = \lambda_2^{Summer} = ... = \lambda_S^{Summer}$	0.000	0.000
Share of $\lambda^{Summer} > 0$	0.213	0.263
Share of $\lambda^{Summer} = 0$	0.512	0.419
Share of $\lambda^{Summer} < 0$	0.275	0.318
Panel B: θ_1^{Summer} and $\widetilde{\theta}_2^{Summer}$		
θ_1^{Summer}	0.243***	0.235***
	(0.004)	(0.003)
$\widetilde{\theta}_2^{Summer}$	0.153***	0.116***
	(0.002)	(0.003)
Panel C: Distribution of Bias in Schools' Value-Add		
Standard Deviation of Unit-Specific Biases	0.619	0.547
Standard Deviation of Spring-to-Spring Value-add	2.448	2.062
Pearson Correlation (Bias, % FRPL)	-0.616	-0.614
Pearson Correlation (Bias, % Minority)	-0.408	-0.421

Table 4: Correlation of Schools' Math and Reading Value-Add and Demographics

	Math λ^{Spring}	Math λ^{Fall}	Reading λ^{Spring}	Reading λ^{Fall}	Percent FRPL	Percent Minority
Math λ^{Spring}	1.00					
Math λ^{Fall}	0.85	1.00				
Reading λ^{Spring}	0.61	0.40	1.00			
Reading λ^{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	-0.56	-0.29	0.70	1.00

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

Panel A: Full Sample

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

Panel B: Bottom Quintile FRPL (Least Poor)

	(Bottom)				(Top)	
	Q_1 λ^{Fall}	Q_2 λ^{Fall}	Q_3 λ^{Fall}	Q_4 λ^{Fall}	Q_5 λ^{Fall}	Total
(Bottom) Q_1 λ^{Spring}	7	1	1	0	0	9
Q_2 λ^{Spring}	9	1	0	0	0	10
Q_3 λ^{Spring}	2	9	10	2	0	23
Q_4 λ^{Spring}	1	10	13	19	2	45
(Top) Q_5 λ^{Spring}	0	1	5	15	47	68
Total	19	22	29	36	49	155

Panel C: Top Quintile FRPL (Most Poor)

	(Bottom)				(Top)	
	Q_1 λ^{Fall}	Q_2 λ^{Fall}	Q_3 λ^{Fall}	Q_4 λ^{Fall}	Q_5 λ^{Fall}	Total
(Bottom) Q_1 λ^{Spring}	39	20	8	1	1	69
Q_2 λ^{Spring}	2	8	13	5	1	29
Q_3 λ^{Spring}	0	2	10	17	4	33
Q_4 λ^{Spring}	0	0	0	5	13	18
(Top) Q_5 λ^{Spring}	0	0	0	0	5	5
Total	41	30	31	28	24	154

Table 6: Transition Matrix for Schools' **Reading** Fall-to-Spring (λ^{Fall}) and Spring-to-Spring (λ^{Spring}) Value-Add

Panel A: Full Sample

	(Bottom)				(Top)	
	Q_1 λ^{Fall}	Q_2 λ^{Fall}	Q_3 λ^{Fall}	Q_4 λ^{Fall}	Q_5 λ^{Fall}	Total
(Bottom) Q_1 λ^{Spring}	110	34	7	2	2	155
Q_2 λ^{Spring}	38	59	47	9	2	155
Q_3 λ^{Spring}	7	48	62	31	7	155
Q_4 λ^{Spring}	0	12	36	69	38	155
(Top) Q_5 λ^{Spring}	0	2	3	44	105	154
Total	155	155	155	155	154	774

Panel B: Bottom Quintile FRPL (Least Poor)

	(Bottom)				(Top)	
	Q_1 λ^{Fall}	Q_2 λ^{Fall}	Q_3 λ^{Fall}	Q_4 λ^{Fall}	Q_5 λ^{Fall}	Total
(Bottom) Q_1 λ^{Spring}	4	1	0	0	0	5
Q_2 λ^{Spring}	3	3	2	0	0	8
Q_3 λ^{Spring}	3	8	11	2	0	24
Q_4 λ^{Spring}	0	8	11	16	5	40
(Top) Q_5 λ^{Spring}	0	1	2	29	46	78
Total	10	21	26	47	51	155

Panel C: Bottom Quintile FRPL (Least Poor)

	(Bottom)				(Top)	
	Q_1 λ^{Fall}	Q_2 λ^{Fall}	Q_3 λ^{Fall}	Q_4 λ^{Fall}	Q_5 λ^{Fall}	Total
(Bottom) Q_1 λ^{Spring}	56	19	5	1	1	82
Q_2 λ^{Spring}	4	11	14	6	2	37
Q_3 λ^{Spring}	0	0	5	8	6	19
Q_4 λ^{Spring}	0	0	1	1	6	8
(Top) Q_5 λ^{Spring}	0	0	0	0	8	8
Total	60	30	25	16	23	154

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

	Quintile of Percent FRPL				
	(Least)				(Most)
Math Value-Add	Q_1	Q_2	Q_3	Q_4	Q_5
Quintile (λ^{Spring}) > Quintile(λ^{Fall})	41.9%	32.9%	26.6%	11.0%	2.6%
Quintile (λ^{Spring}) = Quintile(λ^{Fall})	54.2%	55.5%	59.1%	60.6%	43.5%
Quintile (λ^{Spring}) < Quintile(λ^{Fall})	3.9%	11.6%	14.3%	28.4%	53.9%
	(Least)				(Most)
Reading Value-Add	Q_1	Q_2	Q_3	Q_4	Q_5
Quintile (λ^{Spring}) > Quintile(λ^{Fall})	41.9%	35.5%	27.3%	14.8%	3.2%
Quintile (λ^{Spring}) = Quintile(λ^{Fall})	51.6%	48.4%	51.3%	58.1%	52.6%
Quintile (λ^{Spring}) < Quintile(λ^{Fall})	6.5%	16.1%	21.4%	27.1%	44.2%

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*, 95-135.

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis, 23*(2), 171.

Atteberry, A.& A. McEachin (2015). School's Out: The Role of Summers in Understanding Achievement Disparities. Working Paper.

Baker, G. P. (2000). The Use of Performance Measures in Incentive Contracting. *American Economic Review, 90*(2), 415-420.

Balfanz, R., Legters, N., West, T. C., & Weber, L. M. (2007). Are NCLB's measures, incentives, and improvement strategies the right ones for the nation's low-performing high schools? *American Educational Research Journal, 44*(3), 559-593. doi: 10.3102/0002831207306768

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

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). The impact of assessment and accountability on teacher recruitment and retention: Are there unintended consequences? *Public Finance Review, 36*(1), 88-111.

Castellano, K.E., & Ho, A.D. (2013). Constrasting OLS and quantile regression approaches to student growth percentiles. *Journal of Educational and Behavioral Statistcs, 38*(2), 190-215.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes ? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Charbonneau, ., & Van Ryzin, G. G. (2011). Performance measures and parental satisfaction with New York City schools. *American Review of Public Administration.*

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics, 126,*(4), 1593-1660.

Chetty, R., Friedman, J., & Rockoff, J. (2014a). Measuring the impacts of teachers I: Evaluating the bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632.

Chetty, R., Friedman, J., & Rockoff, J. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633-2679.

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), Holding schools accountable: Performance-based reform in education. Washington, D.C.: The Brookings Institution.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review.*Review of Educational Research, 66*(3), 227.

Dee, T. S., & Jacob, B. A. (2011). The impact of no Child Left Behind on student achievement.*Journal of Policy Analysis and Management*, 418-446

Deming, D.J. (2014). Using school choice lotteries to test measures of school effectiveness. NBER Working Paper 19803.

Dobbie, W., & Fryer, R.G. (Forthcoming). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy.*

Downey, D. B., Von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review, 69*(5), 613.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). Selecting growth measures in school evaluation systems: Should proportionality matter? Forthcoming in *Educational Policy.* doi:10.1177/0895904814557593

Ehlert, M., Kodel, C., Parsons, E., & Podgursky, M. (2013b). 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.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics, 90*(4-5), 837-851.

Figlio, D.N., & Getzler, L.S. (2006). Accountability, ability, and disability: Gaming the system? *Advances in Applied Microeconomics, 14*, 35-49.

Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics, 93*(9-10), 1069-1077.

Figlio, D. N., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. J. Machin & L. Woessmann (Eds.), *Handbooks in Economics: Economics of Education* (Vol. 3, pp. 383-421). North-Holland, The Netherlands: Elsevier.

Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review, 94*(3), 591-604.

Fitzpatrick, M.D., Grissmer, D., & Hastedt, S. (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.

Fuller, S.C., & Ladd, H.F. (2013). School-based accountability and the distribution of teacher quality across grades in elementary schools. *Education Finance and Policy, 8*(4), 528-559.

Gershenson, S. (2013). Do summer time-use gaps vary by socioeconomic status? *American Educational Research Journal, 50*(6), 1219-1248.

Gershenson, S., & Hayes, M.S, (2013). The Implications of Summer Learning Loss for Value-Added Estimates of Teacher Effectiveness. American University School of Public Affairs Research Paper No. 2014-13. Available at SSRN: http://ssrn.com/abstract=2526437

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319), 589-612.

Guarino, C.M., Reckase, M.D., & Wooldridge, J.M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1), 117-156.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Hastings, J.S., & Weinstein, J.M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *Quarterly Journal of Economics, 123*(4), 1373-1414.

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

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization, 7*, 24-52.

Horvath, H. (2015). Classroom assignment policies and implications for teacher value-added estimation. Working Paper.

Jacobsen, R., Saultz, A., & Snyder, J.W. (2013). When accountability strategies collide: Do policy changes that raise accountability standards also erode public satisfaction? *Educational Policy, 27*(2), 360-389.

Jackson, C.K. (2012). Non-cognitive ability, test scores, and teacher quality: Evidence from 9^{th} grade teachers in North Carolina. NBER Working Paper #18624.

Kane, T.J., McCaffrey, D.F., Miller, T., & Staiger, D.O. (2013). Have we identified effective teachers? Validiting measures of effective teaching using random assignment. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T.J., & Staiger, D.O. (2008). Estimating teacher impacts on student achievement: An Experimental Evaluation. NBER Working Paper 14607.

Koedel, C. & Li, J. (Forthcoming). The Efficiency Implications of Using Proportional Evaluations to Shape the Teaching Workforce. *Contemporary Economic Policy.*

Krieg, J. M., & Storer, P. (2006). How much do students matter? Applying the Oaxaca Decomposition to explain determinants of Adequate Yearly Progress. *Contemporary Economic Policy, 24*(4), 563-581. doi: 10.1093/cep/byl003

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*(3), 426-450. doi: 10.1002/pam

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. Educational Researcher, 32(7), 3-13. doi: 10.3102/0013189x032007003

Mathios, A. D. (2000). The impact of mandatory disclosure laws on product choices: An analysis of the salad dressing market. *Journal of Law and Economics, 43*(2), 651-678.

McCaffrey, D.F., Sass, T.R., Lockwood, J.R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

McEachin, A., & Polikoff, M. S. (2012). We are the 5%: Which schools would be held accountable under a proposed revision of the Elementary and Secondary Education Act? *Educational Researcher, 41*(243), 244-251.

Meng, X.L., Rosenthal, R., & Rubin, D. (1992). Comparing correlated correlation coefficFients. *Psychology Bulletin, 111*(1), 172-175.

Mihaly, K., McCaffrey, D.F., Lockwood, J.R., & Sass, T.R. (2010). Centering and reference groups for estimates of fixed effects: Modifications to felsdvreg. *Stata Journal, 10*(1).

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics, 92*(2), 263-283. doi: 10.1162/rest.2010.12318

Papay, J. (2011). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal, 48*(1), 163-193.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37*(1), 7-63.

Polikoff, M.S., McEachin, A., Wrabel, S.L., & Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher, 43*, 45-54.

Reardon, S. F., & Raudenbush, S. (2009). Assumptions of value-added models for estimating school effects.*Education Finance and Policy, 4*(4), 492-519.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics, 92*(5-6), 1394-1415. doi: 10.1016/j.jpubeco.2007.05.003

Rothstein, J. (2014). Revisiting the impacts of teachers. Working Paper.

Rothstein, J. (2011). Teacher quality in educational production: Tracking, decay, and student achievement. Quarterly Journal of Economics, 125(1), 175-214.

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). Grading education: Getting accountability right. Washington, D.C. and New York, N.Y.: Economic Policy Institute and Teachers College Press.

Smith, M. S., & O'Day, J. A. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), The politics of curriculum and testing. New York, NY: Falmer Press.

Todd, P., & Wolpin,K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113*, F3-F33.

Weiss, M.J., & May, H. (2012). A policy analysis of the federal growth model pilot program's measures of school performance: The Florida case. *Education Finance and Policy, 7*(1), 44-73.

Whitehurst, G.J., Chingos, M.M., Lindquist, K.M. (2014). Evaluating teachers with classroom observations: Lessons learned in four districts. Washington, D.C.: Brown Center on Education Policy at Brookings.

Appendix A

In this Appendix, we provide more detail on our value-added model. In order to avoid multicollinearity when estimating a DOLS value-added model, as in When estimating a DOLS value-added model, as in (2), 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 around this arbitrary holdout school. Unless the holdout school is of particular interest, the absolute magnitude of schools' value-add 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 programs in any statistical software program.

It is a two-step process to replicate our value-added model 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)–i.e., the coefficients from (2). Second, use the following formula to generate school value-added estimates from the stored coefficients and independent variables:

$$(\overline{Y_s} - \overline{X}_s\boldsymbol{\beta}) - (\overline{Y.} - \overline{X}.\boldsymbol{\beta}), \tag{A.1}$$

where $\overline{Y_s}$ and $\overline{X_s}$ are school averages of the dependent and independent variables in (2), and $\overline{Y.}$ and $\overline{X.}$ are the unweighted averages of $\overline{Y_s}$ and $\overline{X_s}$, and β is a vector of all of the within-school coefficients from (2).[11] We can see from (A.1) that if any of the variables in

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

$\overline{X}_s$ are correlated with students' summer learning loss, we will get biased estimates of $\boldsymbol{\beta}$, and in turn biased measures of schools' math and reading value-add. To see this more clearly, define $\boldsymbol{\beta^{Short}}$ as the regression coefficients from (2) that does not control for students' summer learning, and define $\boldsymbol{\beta^{Long}}$ as the regression coefficients from (3) if students' summer learning was included in the model. We also know from the omitted variable bias formula, that $\boldsymbol{\beta^{Short}}$=$\boldsymbol{\beta^{Long}}$+$bias(\lambda)$, where, ignoring students' summer learning in the off subject for the moment, $bias(\lambda) = \frac{Cov(X^*_{igst}, SL^*_{igst})}{Var(X^*_{igst})}\theta_1^{Summer}$. Substituting this back into (A.1) and rearraning, we get the following formula for spring-to-spring and fall-to-spring value-add:

$$\lambda^{Spring} = (\overline{Y_s} - \overline{X}_s\boldsymbol{\beta^{Long}}) - (\overline{Y.} - \overline{X}.\boldsymbol{\beta^{Long}}) - (\overline{X}_s - \overline{X}.)bias(\lambda) \tag{A.2a}$$

$$\lambda^{Fall} = (\overline{Y_s} - \overline{X}_s\boldsymbol{\beta^{Long}}) - (\overline{Y.} - \overline{X}.\boldsymbol{\beta^{Long}}) - (\overline{SL}_s - \overline{SL}.)\theta_1^{Summer} \tag{A.2b}$$

If we subtract λ^{Fall} from λ^{Spring}, similar to (4), we get[12]:

$$\lambda^{Spring} - \lambda^{Fall} = \left[(\overline{SL}_s - \overline{SL}.) - (\overline{X}_s - \overline{X}.)\frac{Cov(X^*_s, SL^*_{igst})}{Var(X^*_s)}\right]\theta_1^{Summer}. \tag{A.3}$$

There are three important takeaways from this formulation of bias in schools' value-add from students' summer learning. First, and most obvious, is that students' summer learning does not pose a problem if, conditional on the other independent variables in (2), it does not predict students' spring achievement (e.g., $\theta_1^{Summer} = 0$). Second, we show that students' summer learning loss can take one of two paths (or both) to bias schools' value-add. The first path is through between-school variation in aggregate summer learning $(\overline{SL}_s - \overline{SL}.)$, similar to the auxiliary regressions (5a) and (5b). Even if the covariates in (2) are independent of summer learning, schools' value-add can still be biased

[12] With a little algebra, it can be shown that (A.3) is equivalent to (4) shown in the text. Rearranging (A.3), we get $\left[\left(\overline{SL}_s - X_s\frac{Cov(X^*_{igst}, SL^*_{igst})}{Var(X^*_{igst})}\right) - \left(\overline{SL}. - X.\frac{Cov(X^*_{igst}, SL^*_{igst})}{Var(X^*_{igst})}\right)\right]\theta_1^{Summer}$, which is equivalent to the bias term in (4).

if students' summer learning is differentially distributed across schools. Although if student and school covariates are independent of students' summer learning and summer learning still varies across schools, then much of this variation could be due to differences in instructional practices.

The second path is through the relationship between the covariates in (2) and students' summer learning. Even if $E\left[\overline{SL}_s - \overline{SL}.\right] = 0$, students' summer learning can bias schools' value-add through its relationship with observable student and school characteristics. Granted, it is unlikely that this scenario would pose a serious threat to school value-add, especially since it does not appear to show up in within-school randomized studies (e.g., Kane & Staiger, 2008). We find $Corr\left(\overline{SL}_s - \overline{SL}., (\overline{X}_s - \overline{X}.)\frac{Cov(X_s^*, SL_{igst}^*)}{Var(X_s^*)}\right) = .45$, and the standard deviation of both terms are roughly equal. This suggests that both sources move in the same direction, and, on average, contribute equally to the bias in schools' value-add. Finally, it is clear from (A.3) that the direction of the bias from students' summer learning is complicated by many factors. However, as reported in the main body of the paper, we find a negative correlation between the bias in schools' math and reading value-add and the share of traditionally underserved students in a school.

Appendix B

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. However, beginning in August of 2007, our 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 the Southern state's districts use the same school year start and end dates that is 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 3 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. While 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 extant research suggests that students learn at a linear rate (Fitzpatrick, Grissmer, & Hastedt, 2011), we follow Quin (2014) 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 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 RMSE of 3.35 when we regress students projected scores on their actual scores. We are therefore confident that the projections are not biasing the results presented in this paper.