Working Paper:

The Impact of Summer Learning Loss on Measures of School Performance

Andrew McEachin[1] and Allison Atteberry[2]

State and federal accountability policies require the ability to estimate valid and reliable measures of school impacts on student learning. The typical spring-to-spring testing window used in accountability policies potentially conflates the amount of learning that occurs during the school-year with the learning that occurs during the summer. We use a unique dataset that allows us to explore the extent to which the summer period poses a threat to the validity of the new, more sophisticated growth models used in state accountability systems. The results of this paper have important implications for the design of future accountability policies.

[1]North Carolina State University
[2]University of Virginia

Updated May 2014

EdPolicyWorks, University of Virginia
PO Box 400879
Charlottesville, VA 22904

McEachin, A. & Atteberry, A. (2014) The Impact of Summer Learning Loss on Measures of School Performance. *EdPolicyWorks Working Paper Series, No. 26.* Retrieved from:
http://curry.virginia.edu/uploads/resourceLibrary/26_McEachin_Summer_Learning_Loss.pdf

Acknowledgements: : 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 Parson 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. The standard caveats apply.

THE IMPACT OF SUMMER LEARNING LOSS ON MEASURES OF SCHOOL PERFORMANCE

Andrew McEachin & Allison Atteberry

Introduction

One of the most prominent debates in education policy today is how to design state and federal 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 the school's influence and the myriad external factors that also contribute but are outside the school's purview. Policy makers at both the federal and state level are in the midst of experimenting with new approaches to this accountability challenge. At the same time, researchers are continuously developing new statistical techniques—what Castellano and Ho (2012) call aggregate-level conditional status metrics (ACSMs) models—that compare the aggregate performance of schools after conditioning on student and school characteristics that are beyond the control of educators and administrators (Castellano & Ho, 2013; Ehlert, Koedel, Parsons, & Podgursky, 2013; Ladd & Walsh, 2002; Reardon & Raudenbush, 2009; Todd & Wolpin, 2003).

The validity of these ACSMs as a measure of school performance rests on a number of assumptions, many of which have been explicated and probed in existing work (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 use of spring-to-spring achievement data in most accountability policies and measures of school performance therefore implicitly assumes that the amount of learning lost/gained over the summer is either virtually the same across students, randomly distributed among students, or is captured by other observable student characteristics. In short, the typical 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 growing research on the impact of the summer period on student summer learning rates suggests that the amount of learning gained/lost over the summer is not uniform across students from various backgrounds. 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; Authors, 2014; Downey, Von Hippel, & Broh, 2004; Gershenson & Hayes, 2013)—

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

the negative impact of summer on lower socioeconomic students is oftern referred to as "summer setback" or "summer learning loss." The role of the differential summer setback in estimating measures of school quality is further complicated by the systemic sorting of students to schools based on student social and economic characteristics.

Starting with the Federal Growth Pilot program in 2005 (Weiss & May, 2012), which allowed states to incorporate simple growth-to-proficiency models into their implementation of NCLB, and now with the ESEA waiver program (Polikoff, McEachin, Wrable, & Duque, 2014), states have started to hold schools accountable not only for students' achievement levels but also their achievement growth over time using some form of ACSMs. It is hoped that the use of student achievement growth will better capture the portion of student achievement attributable to schools' policies and practices, as well as reduce the link between schools' accountability standings and outside of school factors (e.g., student poverty), eliminating many of the concerns associated with older measures based on achievement levels (Elhert et al, 2013a). However, no research to date has examined the potential for "summer setback" to bias commonly used ACSMs in school accountability policies.

In this paper we use a unique dataset that contains both fall and spring test scores for students in grades 3 through 8 from a Southern state to evaluate the impact of summer setback on growth models commonly used in state accountability systems to estimate school quality. We start by using an omitted variable bias framework to anticipate the conditions under which failing to address summer learning loss would bias estimated school growth measures. We further explore the extent to which those conditions seem to be met. We estimate two variants of ACSMs—student growth percentiles (SGPs) and value-added models (VAMs)—to examine the systematic differences in schools' math and reading performance measures according to the testing window used: spring-to-spring and fall-to-spring. Finally, we take on a policy lens and consider the extent to which specific schools are ranked quite differently depending on whether fall data is used to account for summer learning loss. When schools are ranked lower using only the traditional spring-to-spring data than they are when fall data is also used, we are concerned that they are being unfairly penalized for differential summer learning loss. We are particularly concerned if disadvantaged schools are the ones with the greatest bias due to the oversight of summer learning loss. Specifically, we ask the following questions:

Q1: Is there evidence that summer learning loss may be distributed across students and schools in systematic ways that may bias school-level ACSMs?

Q2: How correlated are the schools' ACSMs to each other and school demographics?

Q3: How does schools' relative ranking change between a spring-to-spring and fall-to-spring test timeline?

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 accountability performance measures. We find that schools serving more disadvantaged students are more likely to be identified among the lowest-performing when ACSMs rely on the traditional spring-to-spring test timeline rather than the more accurate fall-to-spring timeline. The rest of the paper proceeds as follows: We review the relevant ACSM and summer learning literature. We then discuss the unique data set used to answer our research questions, followed by a description of the methodological approach. We then present the results and close with a discussion of conclusions and limitations.

Relevant Literature

Accountability

School accountability policies have been implemented to solve two basic problems in public education: principal-agent problem and information asymmetry (Baker, 1992; Figlio & Kenny, 2009; Figlio & Lucas, 2004; Holmstrom & Milgrom, 1991; Ladd & Zelli, 2002; Prendergast, 1999; Mathios, 2000; Reinstein & Snyder, 2005). The former assumes that the use of performance incentives (e.g., rewards and sanctions) will better align educators' behaviors with local, state, or federal standards (Prendergast, 1999; Holmstrom & Milgrom, 1991; Smith & O'day, 1991). The latter problem is mitigated by infusing the educational marketplace with information about the effect of schools on students' achievement and other outcomes (Charbonneau & Van Ryzin, 2011; Figlio & Loeb, 2011; Jacobsen, Snyder, & Saultz, 2013; Rothstein et al., 2008). In both cases, performance is generally defined in terms of students' achievement on test scores, which presumes that tests scores, despite not capturing every skill deemed important by society, are strongly related to students' future success (e.g., labor market earnings) (Chetty, Friedman, & Rockoff, 2012).

Beginning in 2002, *No Child Left Behind* (*NCLB*) required every state to implement test-based accountability systems. According to federal guidelines, each state articulated how schools would experience sanctions if their performance on state-selected standardized tests consistently falls below

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

a specified level. This federal directive reflects a general increase in the use of test scores as a barometer of school quality (Harris & Herrington, 2006). Indeed, the 2001 passage of *NCLB* signaled that test-based accountability is now a cornerstone of the federal government's role in schools (DN Harris, 2009).

Now over a decade later, research has shown that performance-based school accountability policies, including *NCLB*, generally raise students' average achievement (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005; Rouse, Hannaway, Goldhaber, & Figlio, 2013). At the same time, research also reveals that these systems are prone to a number of sources of bias and unintended consequences. These problems arise because the real-world policies naturally deviate from the idealized theory of action articulated above. Of course no policy is perfect, however it is important to document when design flaws undermine the central behavioral mechanism, or worse yet exacerbate the very inequalities the policy was designed to address.

One key assumption of accountability theory is that it is possible to estimate measures of school performance that accurately reflect actors' efforts. If these measures are too noisy, too hard to move, or unduly influenced by factors outside the actors' control, the incentives to align behaviors with expectation break down, and unintended consequences may emerge. In its current form, *NCLB* requires schools to make 'adequate yearly progress' (AYP) in terms of the percentage of students who demonstrate proficiency on statewide exams. This approach does not take into account that some schools tend to receive students that are initially farther from the state's definition of proficiency. These pre-existing differences in performance arise in large part due to non-school factors that affect how students perform on tests, in combination with the nonrandom sorting of students into schools. Since traditional accountability formulae do not take these differences into account, schools that serve students far from statewide targets are less likely to make AYP and more likely to be sanctioned.

Accountability systems assume that many parties will respond to the signal provided by performance measures. Research shows that families, for instance, respond to the information produced by accountability systems in many ways, including decisions about where to buy a house (Figlio & Lucas, 2004), whether to donate to the public school system (Figlio & Kenny, 2009), or whether to participate in inter-district transfer programs (Hastings & Weinstein, 2008). Educators also appear to respond to accountability pressures in a number of ways, such as reallocating teachers across tested grades and subjects (Fuller & Ladd, 2013), narrowing the curriculum to tested subjects

and students on the margin of passing the state proficiency threshold (Neal & Schanzenbach, 2010), and differential treatment of students more likely to fail achievement tests (Figlio, 2006; Figlio & Getzler, 2006). The information generated from school accountability policies also influences the teacher labor market, encouraging teachers to seek out more stable positions in schools that are not in danger of failure.

It is hotly debated whether these behavioral responses on the part of families, school leaders, and teachers are beneficial to schools or not. However, regardless of one's take on these policy mechanisms, it is clear that *none* of these behavioral responses can be beneficial if the accountability system identifies the wrong schools as highly effective or ineffective. For instance, if school performance measures over-identify and penalize schools serving larger shares of traditionally under-served students, effective teachers may be incentivized to avoid working with the most-disadvantaged students (Balfanz, Legters, West, & Weber, 2007; Boyd, Lankford, Loeb, & Wyckoff, 2008; Krieg & Storer, 2006).

In an effort to mitigate the unintended consequences that arise due to flawed school performance measures, the federal department of education has incentivized states (e.g., the ESEA waiver program) to move away from overly simplistic measures of school performance such as AYP or percent proficient in favor of measures of student achievement growth, or growth models. At its core, any growth model (herein, ACSMs) estimates the amount of learning a student gained in one school-year, and then aggregates the learning within a school to generate a school-level performance measure (Castellano & Ho, 2013). The performance of a given school is then compared to a referent distribution, often conditional on student and school characteristics. The inclusion of students' aggregate growth in accountability policies opens up the possibility that schools with low levels performance may be acknowledged if their students make above-average improvements during the school-year.

There are two prominent methods for estimating measures of student achievement growth: Student Growth Percentiles (SGP) and value-added models (discussed in more detail in the methods section). The former is used by at least 27 states, while the latter is prominent in the research literature and teacher accountability. Though there is hope that the use of more nuanced models can mitigate some of the unintended consequences that arise from using overly-simplistic school performance measures, concerns remain about potential bias in ACSMs. Researchers have begun to document the sensitivity of growth-based ACSMs to model specification, measurement error, and

year-to-year instability (Elhert et al, 2013b; Goldhaber & Hansen, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Papay, 2011). To date, no study has evaluated the impact of conflating the summer period with the school year on ACSMs. Under certain conditions, we hypothesize that the move toward improved school performance measures like ACSMs may not address the bias introduced by relying on once-per-year test score data to estimate school impacts.

Summer Setback

There is growing evidence that disadvantaged children have fewer learning opportunities during the summer months than their advantaged counterparts. Heyns (1978) separated students' cognitive gains during the summer versus gains during the school-year, because she acknowledged that, by definition, summer learning reflects only the influence of non-school factors. Analyzing a sample of approximately 3,000 fifth and sixth graders in Atlanta, Heyns 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, Entwisle and Alexander (1992, 1994) analyzed children's fall and spring test scores in Baltimore. In their sample, Entwisle and Alexander found that both socioeconomic and race gaps in reading skills grew at faster rates during the summer. More recently, Downy, Von Hippel, & Broh (2004) have even argued that seasonal comparison research suggests that the socioeconomic and racial/ethnic gaps in reading and math skills which garner so much attention in today's conversation about U.S. public schools are not in fact the product of an unequal school systems, but instead are widened primarily during the summer.

However little is known about precisely what causes students of different socioeconomic and demographic backgrounds to experiences summers so differently, though research has suggested that income differences could be related to differences in opportunities to practice and learn over summer, with more books and reading opportunities available for middle-class children (Cooper et al., 1996; Downey et al., 2004; Heyns, 1978). Furthermore, Gershenson (2013) found that low-income students watch two more hours of TV per day during the summer than students from wealthier backgrounds. Because the sources of differences is summer learning are unobserved, value-added models that conflate the summer period with the school-year will inappropriately blame teachers with disadvantaged students for this summer decay while the teacher of the advantaged student will be credited for gains they did not foment (see discussion in Scherrer, 2011).

To date, little attention has been paid to the intersection between the summer learning loss literature and ACSMs. This is likely due to the fact that spring-to-spring test timing is a ubiquitous

aspect of statewide testing systems and therefore few opportunities exist to question this aspect of the models. Three other research projects that we are aware of have touched on this subject (Gershenson, & Hayes, 2013; Papay, 2010; Downy, von Hippel, and Hughes, 2008). In his paper on how teacher value-added scores might be sensitive to using different tests, Papay (2010) uses data from a large Northeast district to examine the agreement between value-added models across three different tests given in that district. Though it is not the main point of the paper, Papay notes that one of his three tests is given both in the fall and spring of each school-year, and he runs teacher value-added models that use a spring-spring and fall-spring value-added timelines. He finds that the correlation between spring-spring and fall-spring is between 0.66 or 0.71 depending on the sample. Gershsenson and Hayes (2013), using data from the Early Childhood Longitudinal Study of 1998-99 (ELCS), finds similar correlations among spring-to-spring and fall-to-spring value-teacher added measures. The authors also find that the relationship between observable teacher inputs and student outcomes differ when student growth is measured on different test-timelines. Lastly, Downy, von Hippel, and Hughes (2008) also use ECLS data to estimate random-effect growth models, and they find that schools serving larger shares of low-income students are more likely to be in the bottom of the performance distribution when school performance measures do not account for the summer period.

The three studies suggest that ignoring the summer period will disproportionately affect schools serving larger shares of minority and low-income students under traditional accountability regimes. However, the studies also raise important unanswered questions. The Papay (2010) and Gershenson and Hayes (2013) papers do not investigate the relationship between the spring-to-spring and fall-to-spring value-added discordance and student and class demographics. For example, if teachers with more difficult to educate students have larger fall-to-to spring value-add than spring-to-spring, and vice versa for teachers with easier to educate students, then their studies raise important equity issues related to the design of the modal accountability system. The three studies also have important data limitations. The studies either rely on a few years of data within one urban district, or follow one cohort of nationally representative students over time. Downy, Hippel, and Hughes (2008), the only study to evaluate this phenomenon at the school-level, do not use ACSMs that are commonly used for school accountability policies or in the education policy literature.

The results of our paper address the gaps in the interrelated accountability, ACSM, and summer learning literatures in three ways. First, we utilize a state-wide panel of student achievement

data from grades 3 through 8 over a five-year period. Instead of relying on the summer period between kindergarten and first grade (Gershenson & Hayes, 2013; Downy, Hippel, & Hughes, 2008), the grade span used in this study is more representative of the grades typically included in high-stakes accountability policies. Second, we are the first to evaluate the impact of summer setback on ACSMs used in state accountability policies and the research literature. Lastly, we not only examine whether summer setback leads to misclassifications in ACSMs, but also the types of schools that are most affected by this phenomenon.

Data

The data for this study is from the North West Evaluation Association's (NWEA) Measures of Academic Progress (MAP) assessment. The MAP is a computer adaptive test given in math, reading, science, and social studies in over 40 states in the U.S. To ensure that the MAP scores provide a valid measure of students' knowledge, the NWEA aligns the MAP items with the specific state standards. The MAP is also scored using a vertical and interval scale, which the NWEA calls the RIT scale. The vertical scale allows comparisons of student learning across grades and over time, while the interval scale ensures that a unit increase in a student's score represents the same learning gain across the entire distribution.[i] In sum, the MAP assessment has appropriate measurement properties for a study of school value-added models and summer setback.

The data for our study comes from a Southern state that administered the MAP assessment in the fall and spring for all students in grades 2 through 8 for 2007-8 through 2009-10 school-years. The MAP assessment was used as a supplementary tool to aid schools' in improving their instruction and meeting students' needs, not as the high-stakes test of record. While the low-stakes nature of the MAP assessment in this state reduces the validity threat of teachers focusing their instruction specifically to tested material, it raises another validity threat. It may be the case that although the MAP assessment covers the standard teachers are expected to cover during the school-year, the low-stakes nature of the test does not incentivize students to try their best. However, it is unlikely that educators and state administrators would continue to request the NWEA to administer the test twice a year to all students in grades 3 through 8 if the test was not taken seriously.

Our dataset includes student- and school-level files that are longitudinally matched over time. The student-level file includes basic demographic information, such as the students race and gender, their math and reading scores, the measurement error associated with their math and reading

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

scores, grade of enrollment, the date of test administration, and fall and spring school IDs. Noticeably, 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 program, and participates in special education. The school-level data file is provided by the Common Core of Data through the NWEA. This data includes the typical set of school-level characteristics, including the percent of FRPL students within the school. The student- and school-level descriptives for the 2009-10 school-year are provided in Table 1.

Analytic Samples

We use SGPs and value-added models to estimate three-year average school math and reading effects. In order to make proper comparisons across each of the testing windows, it becomes important to carefully define the analytic sample. 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 2006-07 school-year, this means the first year we can generate a school performance measure using the spring-to-spring test timeline is 2007-08—using the 2006-07 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. To remain consistent between test timelines we also use grades 3 through 8 in all years 2007-8 through 2009-10. To remain consistent between test timelines, we estimate three-year school SGPs and value-add for the spring-to-spring and fall-to-spring test timeline using the 2007-8, 2008-9, and 2009-10 school-years. Our analytic sample includes students in grades 3 through 8 during the 2007-8 through 2009-10 school-years with approximately 45,000 students per grade per year.

Methods

The goal of our analyses in this paper is twofold. We first estimate a series of school performance measures to evaluate the impact of students' differential summer setback on schools' accountability standing. In the second part, we are specifically interested in the impact of switching from a spring-to-spring test timeline to a fall-to-spring test timeline, holding other aspects of the model constant. In this section, we delineate the methods used to accomplish these tasks.

Estimating School Performance Measures

Student growth percentiles. Student growth percentiles—SGPs—are the most commonly used example of an ACSMs in school accountability systems. SGPs are the conditional median

percentile rank of students' current achievement based on their prior achievement histories for a given school. Students are given an SGP of 1 to 99, representing the student's achievement percentile rank for a given year relative to students with similar achievement histories. To use the SGP in school accountability policies, states then use the median SGP within a school to generate an aggregate performance measure. For example, a school with a median SGP of 65 means that the median student in the school had a percentile rank of current achievement at the 65th percentile.

SGPs are typically estimated using a complex B-spline quantile regression method (Betenbenner, 2011). However, researchers have found that a simple linear model produces SGPs that are highly correlated (r~.96 to .99) to the more complicated quantile regression method (Elhert et al, 2013a; Castellano & Ho, 2013). Conceptually, the linear SGPs are constructed in three steps. First students' achievement in the current year is regressed on as many years of prior achievement as possible (usually estimated separately by subject, grade, and year). The residual of this linear regression represents the difference between students' predicted achievement in the current year based on their prior achievement histories and their actual achievement. Second, the student-level residual from the first regression is converted into a percentile rank. Third, the residuals from the same subject (e.g., Reading) are pooled across grades and years and the within-school median of the percentile is the school's median SGP.

The sparse nature of the SGPs leaves the model open to many sources of potential omitted variable bias if there are factors that are related to students' current achievement that are not captured by students' prior achievement. For example, extant research and our own findings below find a strong negative correlation between schools' SGP and the percent of students living in poverty (Elhert et al, 2013a). Furthermore, in a simulation study of SGPs and other ACSMs used in teacher and school accountability policies, Guarino, Reckase, Stacy, and Wooldridge (2014) find that SGPs do less well than other models in reproducing true teacher effects when students are non-randomly assigned to classes. Given that students are not randomly assigned to schools, it is likely that SGPs are less likely to reproduce true school effects as well. It is also unclear how summer setback will affect the use of SGPs in school accountability policies.

Instead of using the B-spline quintile regression package in R to estimate our median SGPs, we use an OLS approach described below since the OLS approach is more conceptually similar to the value-added modeling approach used in this paper, and since it allows us to control for differential amounts of instructional days across students, When we compare our OLS approach to

the results from the package in R we obtained nearly identical results with correlations around .96 to .99.

We use the below OLS regression model as our baseline method for calculating students' and schools' SGP (estimated separately by subject, grade, and year):

$$(1) \quad Y_{igst}^{Spr} = \beta_0 + \theta_1 Y_{igs,t-1}^{Spr} + \sum_{p=2}^{4} (\theta_p Y_{igs,t-p}^{S} + \gamma_p^s * Missing_{igs,t-p}) + \sum_{j=0}^{3} \tau_j days_{igs,t-j} + \varepsilon_{igst}$$

In our basic SGP model in Equation (1), the outcome, Y_{igst}^{Spr} , is the current spring achievement score of student i in grade g in school s in school-year t, and it is modeled as a linear additive function of the student's spring achievement in the same subject in prior years t-1 (as far back as school-year t-4). Note that each student in the NWEA data actually possesses two scores in each school-year t—that is, *both* a spring score Y_{igst}^{Spr} and a fall score Y_{igst}^{Fall}— however in model (1) we do not include any fall scores in order to simulate the traditional spring-to-spring test score data used in virtually all statewide testing systems. Because we observe specific test dates, we can also control for "$days_{igst}$", the number of days since each student's previous (spring) test.[ii] We stack the residuals across grades and years from model (1) and use the within-school median as each school's SGP measure.

Unlike value-added models that control for the contemporaneous relationship student, peer, and school inputs, as well as a student's prior achievement, model (1) assumes a student's cumulative achievement history captures the effect of these input on current achievement levels.[iii] Model (1) is a sparse cumulative value-added model described in Todd and Wolpin (2003) with three important assumptions: (a) Contemporaneous student, peer, and school inputs do not affect students' achievement level; (b) time-varying and fixed student, family, peer, and school characteristics are captured by the prior achievement histories; (c) Instead of assuming a constant geometric decay in prior students' achievement, model (1) directly estimates the decay of each of the prior periods available to the researcher. If either assumption (a) or (b) fails, a school's SGP measure will be potentially conflated with factors related to a student's achievement that is not within the school's control, including the amount of learning/less that occurs over the summer.

In order to assess the impact of summer learning loss on school SGPs, we estimate four versions of the SGP model (1):

- Model (1), as shown;
- Model (1a): add the lagged off-subject achievement (e.g., reading for the math SGP);
- Model (1b): add both the fall and spring lagged same subject achievement;
- Model (1c): add fall and spring lagged achievement for both the same and off-subject.

We estimate each model separately by subject, grade, and year for the 2009, 2010, and 2011 school-years. Even though states typically only control for the prior achievement histories in the subject of interest, we include the off-subject specifications in SGP model (1a) to account for between student differences that the subject of interest does not capture. Model (1b) adds the lagged fall scores to the original SGP model. Finally, in Model (1c), we incorporate all available achievement data from both previous falls and springs and in both subjects. In each year, we use as many prior achievements as possible (e.g., up to 2 for 2009 and 4 for 2011), and we stack the residual from the three years to create one SGP. It is important to note that most states only use the SGP from one year in their accountability models. Since our three-year SGP uses more data, it is possible our results capture the lower bound of the true summer loss problem with a spring-to-spring SGP.

Value-added models. A second approach to measuring a school's contribution to student achievement growth is the set of models generally referred to as "value-added." By taking into account the pre-existing student characteristics—such as students' race/ethnicity, free and reduced price lunch program (FRPL) status, and so on—as well as students' prior achievement, the value added by schools can, in theory, be statistically disentangled from the characteristics of students that are outside the school's control. The appeal of this approach is obvious: VAMs have the potential to remove the effects of teachers and schools from the effects of students family and community influence, thereby holding teachers and schools accountable for the amount of, or changes in, student achievement within their control (McCaffrey, 2003).

The most common value-added specification is an education production function dynamic OLS (DOLS) model that regresses a students' current achievement on her prior achievement and teacher and/or school fixed-effects (depending on the level of the accountability policy), and potentially vectors of student and school control variables, also known as a dynamic OLS (DOLS) value-added model (Guarino, Reckase, & Wooldridge, 2012; Todd & Wolpin, 2003). Although questions have been raised about the ability of this specification to produce unbiased effects of schools on students' achievement (Rothstein, 2011), researchers have used the DOLS specification

to replicate school effects within experimental studies (Chetty, Friedman, & Rockoff, 2013; Kane et al, 2013; Kane & Staiger, 2008; Deming, 2014).

To estimate school value-added measures, we start with a common DOLS model. Importantly, the value-added model also includes a set of time-invariant and time-varying student demographic characteristics, grade fixed effects, time-varying school characteristics, and school fixed effects (Guarino, Reckase, & Wooldridge, 2012; Todd & Wolpin, 2003):

$$(2) \qquad Y_{igst}^{Spr} = \theta_1 Y_{igs,t-1}^{Spr} + \theta_2 \tilde{Y}_{igs,t-1}^{Spr} + \boldsymbol{X_{igst}\beta_1} + \boldsymbol{Z_{st}\beta_2} + \delta_s + e_{igst}$$

The outcome, Y_{igst}^{Spr} , is the spring MAP achievement score for student i in grade g school s in school-year t. This is modeled as a linear function of spring achievement in the prior school-year t-1 in both the same and off-subject, $Y_{igs,t-1}^{Spr}$ and $\tilde{Y}_{igs,t-1}^{Spr}$, respectively. We herein refer to Model (2) as the "spring-to-spring" version of the value-added model, because it predicts the current spring score using the prior spring scores. This value-added model also includes a vector of student demographic characteristics ($\boldsymbol{X_{igst}}$) including race, an indicator for whether the student made non-structural change of schools between school-years, an indicator for whether students changed schools within the school-year, the number of days of instruction between test administrations[iv], and grade level fixed effects; school level aggregates of the student demographic characteristics ($\boldsymbol{Z_{st}}$) as well as the percent of FRPL students in the school and the natural log of enrollment; a vector of school fixed-effects (δ_s) which indicates the school to which the student is exposed in the given year; and an idiosyncratic student-level error term, e_{igst}. The key parameter of interest is δ_s, which captures the average achievement of a school's students over the three year panel, conditional on their prior reading and math skill and student and school characteristics. We run the model separately for math and reading.

In order to assess the importance of differentiating between the school-year and summer, we make a key change to model (2)—that is, the definition of the prior test score timing on the right-hand side of the equation. In order to assess the impact of summer learning loss on school value-added measures, we estimate two versions of model (2):

- Model (2), as shown (i.e., the "spring-to-spring" specification). Here, we use the students' test scores from the prior spring as right-hand-side variables, as is typical in the majority of value-added applications. We make no use of the fall test score data;

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

- Model (2a): We replace the prior year spring test scores on the right-hand side of the equation with the corresponding fall scores from the current year (i.e., the "fall-to-spring" specification).

$$(2a) \qquad Y_{igst}^{Spr} = \theta_1 Y_{igst}^{Fall} + \theta_2 \tilde{Y}_{igst}^{Fall} + \boldsymbol{X_{igst}\beta_1} + \boldsymbol{Z_{st}\beta_2} + \delta_s + e_{igst}$$

The fall-to-spring specification (2a) removes the summer portion of a students' learning from the estimation of the value-added model, effectively holding schools accountable for the school-year portion of student achievement.

Omitted Variable Bias

Given the structure of the modal assessment system in the US and general absence of fall testing, spring-to-spring SGPs (models (1)) and value-added models (model (2)) are the most common specifications in real-world scenarios. We are concerned, however, that estimates of school quality generated by Models (1) and (2) may be biased because the model conflates school-year learning with summer learning. Take the SGP for example; there are three conditions where the summer learning present in the student-level error term will not pose a problem for the use of SGPs in school accountability policies. The first is that the amount of learning loss over the summer is fully explained by a student's prior achievement level. Since achievement levels are correlated with poverty, and poverty is correlated with summer learning loss, it is possible that a student's prior spring achievement level explains at least a portion of the summer learning loss. However, it is unlikely that it explains *all* of it. Second, if students are randomly assigned to schools then although the summer portion of the error term will not equal zero for each student, it will equal zero on average for each school. Lastly, if we believe that schools should be fully accountable for students' summer behaviors. If these assumptions do not hold, then schools serving students that experience summer learning loss, as opposed to gains, will have a downwardly biased SGP, and vice versa. The value-added specification in Model (2) relaxes the SGP assumptions by including student- and school-level demographic variables in the estimation of school-level performance measures, and reduces the likelihood that the summer learning is a potential source of omitted variable bias. The unique nature of our assessment data allows us to directly examine the potential for the summer period to serve as a source of bias in the estimate of school performance measures.

Even conditional on student and school covariates, Model (2) may produce biased school quality measures in that the achievement gain/loss that occurs in the summer between spring of last school-year (spring of t-1) and fall of the present school-year (fall of t) is inherently part of the

estimated effect of each school, $\hat{\delta}_s$. In order to explore this concern in more depth, we define a new variable, ΔY_{igst}^{Sum}, which is the summer achievement gain/loss for a given student in the summer preceding school-year t, simply calculated by $\left(Y_{igs,t}^{Fall} - Y_{igs,t-1}^{Spr}\right)$.[v]

We wish to directly examine whether the conditions for omitted variable bias to be present in the current scenario. To do so, we conceive of the school-fixed effects from Equation (2) as imperfect estimates of the true impact of schools on student achievement, and the impact of (non-school) summer learning as a potential omitted variable. Ignoring other sources of omitted variable bias, the probability limit of the estimated school fixed-effects, $\hat{\delta}_s$, from Equation (2) yields the well-known omitted variable bias formula[vi]:

$$(3) \qquad \hat{\delta}_s \xrightarrow{p} \delta_s + \frac{Cov(\ddot{\delta}_s, \Delta Y_{igst}^{Sum})}{Var(\ddot{\delta}_s)} \hat{\theta}_3$$

In order to estimate the first term of the bias product, $\frac{Cov(\ddot{\delta}_s, \Delta Y_{igst}^{Sum})}{Var(\ddot{\delta}_s)}$, we modify the spring-to-spring value-added model shown in Equation (2) by replacing the outcome (formerly current spring test scores) with ΔY_{igst}^{Sum} :

$$(4) \qquad \Delta Y_{igst}^{Sum} = \theta_1 Y_{igs,t-1}^{Spr} + \theta_2 \tilde{Y}_{igs,t-1}^{Spr} + \boldsymbol{X_{igst}\beta_1} + \boldsymbol{Z_{st}\beta_2} + \ddot{\delta}_s + \alpha_g + e_{igst}$$

Here, the estimated school fixed-effects, $\frac{Cov(\ddot{\delta}_s, \Delta Y_{igst}^{Sum})}{Var(\ddot{\delta}_s)} = \hat{\ddot{\delta}}_s$, estimated in equation (4) captures the average amount of summer learning within school *s* conditional on student and school covariates. We estimate $\hat{\ddot{\delta}}_s$ using a sum-to-zero constraint so each fixed-effect represents the amount of summer learning within a school relative to the sample average, which is set to zero. A negative value for a given school indicates that, on average, the students in school s experience learning *losses* over the summer, and vice versa. If there is no covariance, then the bias due to summers shrinks to zero. Since the sum-to-zero constraint does not allow us to estimate the mean bias for the sample, we instead conduct a joint hypothesis test to see if the fixed effects $\hat{\ddot{\delta}}_s$ are all equal to zero.

In order to estimate the second term in the omitted variable bias product, $\hat{\theta}_3$, we again modify Equation (2) by adding in the variable ΔY_{igst}^{Sum} as an additional independent variable. The coefficient, θ_3, captures the relationship between summer learning and current spring achievement levels, holding other factors constant.

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

$$(5) \qquad Y_{igst}^{Spr} = \theta_1 Y_{igs,t-1}^{Spr} + \theta_2 \tilde{Y}_{igs,t-1}^{Spr} + \theta_3 \Delta Y_{igst}^{Sum} + \boldsymbol{X_{igst}\beta_1} + \boldsymbol{Z_{st}\beta_2} + \delta_s + \alpha_{gt} + e_{igst}$$

We assume that this term is positive; students with higher levels of learning over the summer also have higher levels at the end of the school-year. The bias presented in (3) is particularly policy relevant if the correlation between the estimated bias and the share of traditionally under-served students within a school is negative, effectively penalizing schools for educating students from difficult backgrounds.

The question becomes: Under what conditions is the bias term equal to zero. One condition is that, on average within a school, students do not experience learning loss (that is, $\Delta Y_{igst}^{Sum} = 0$). In this case, $\frac{Cov(\ddot{\delta}_s, \Delta Y_{igst}^{Sum})}{Var(\ddot{\delta}_s)}$ also equals 0. Given what we know about summer learning loss and student sorting, this is unlikely to occur. Second, summer learning loss is captured by students prior spring achievement and/or other student and school demographics ($\hat{\theta}_3 = 0$). This is also unlikely to occur. However, we do see from (3) that the amount of bias in a school's fixed-effect estimate, ignoring other sources of bias, is weighted by the magnitude of $\hat{\theta}_3$. The structure of our data set allows us to estimate these parameters to measure the potential bias in schools' value-add from students' summer learning. Although we cannot estimate a mean bias for our sample, we are able to correlate schools' bias from (3) with schools' demographic characteristics.

Both the spring-to-spring (Equation 2) and fall-to-spring (Equation 2a) value-added specifications are designed to estimate a school's effect in the year *t*, and the effect is anchored to the school the student attended in the spring of time t. The switch from a fall-to-spring test timeline in Model (2a) will remove the bias in Model (2) and reduce the relationship between aggregate school performance and student demographics. It is unlikely that the fall-to-spring timeline will completely nullify the correlation, as other factors are both related to student demographics and school quality (e.g., the teacher labor market).

Results

Q1**: Is there evidence that summer learning loss may be distributed across students and schools in systematic ways that may bias school-level ACSMs?**

We start our investigation into the potential problems summer setback causes for school accountability policies by estimating the two coefficients of interest in the omitted variable bias

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

formula presented in Equation (3): The first is the term, $\frac{Cov(\ddot{\delta}_s, \Delta Y_{igst}^{Sum})}{Var(\ddot{\delta}_s)}$, which we recover from model (4), and the second term is $\hat{\theta}_3$, which we estimate using equation (5) .

Recall that the coefficients $\hat{\ddot{\delta}}_s$ are the estimated school fixed-effects from model (4). These estimated fixed-effects capture the average amount of summer learning loss within a school after partialing out student and school characteristics. We run an F-test on the joint-hypothesis that $(\ddot{\delta}_s = 0, \forall s = 1, \ldots, S)$; a failure to reject this hypothesis indicates that there is significant variation in summer learning loss across schools, conditional on student and school covariates. As shown in Table 2 we reject the null hypothesis that the fixed-effects are all equal to zero for both math and reading. This result indicates that even conditional on a vector of student and school characteristics, there is non-zero variation in students' summer learning across schools.

The term, $\hat{\theta}_3$, is the coefficient on summer learning loss variable, ΔY_{igst}^{Sum}, included in model (5). It captures the influence of summer learning on students' current spring achievement; we would expect this term to be positive. Given that the average summer learning loss varies across schools, if $\hat{\theta}_3$ is statistically significant and positive, then the value-added model (2) suffers from omitted variable bias due to students' summer learning experiences. The estimated coefficients for math and reading, shown in Table 2 are 0.520 and 0.413, respectively. Even conditional on a host of student and school characteristics, prior achievement, and school fixed-effects, students' summer still influence their spring achievement levels. A one standard deviation decrease in math and reading summer learning, 7.5 and 8.5 RIT points respectively, is associated with an approximately 3.5 RIT decrease in a students' math or reading spring MAP assessment. This decrease is roughly 20 percent of a standard deviation on the spring assessment. To estimate the bias for each school we take the product of each school's fixed-effect from model (4), $\hat{\ddot{\delta}}_s$, and the coefficient for the summer learning in model (5), $\hat{\theta}_3$.

The last three rows of Table 2 provide descriptive information about the implications of the bias in schools' math and reading value-add. The standard deviation of the bias is 0.735 and 0.617 points on the math and reading RIT scale, respectively. To put this in perspective, the standard deviation of the RIT scale for either subject in a given test administrator (e.g., Math in Fall 2009-2010) is 15. Furthermore, we find the standard deviation of schools' math or reading value-add is approximately 1.5 RIT points. The evidence thus far suggests that the traditional spring-to-spring

test timeline does not adequately account for the summer period, and school performance measures (e.g., value-add) generated from this timeline are prone to non-trivial bias. In our case, the standard deviation of the bias is roughly 50 percent of the standard deviation of the estimated school effects shown later in Table 3.

The last two rows in Table 2 evaluate whether the bias in spring-to-spring value-added models differentially affects certain types of schools. The correlation between schools' bias in math value-add and their share of FRPL and/or minority students is statistically significant and moderately negative. The correlation is more substantial for the correlation between reading value-add and the percent of FRPL students in a school (r= -0.5). In Figure 1 we plot the kernel density of the bias separately by tertiles of schools' percent of FRPL students. The conditional median of the math bias is relatively similar across the FRPL tertiles, with more pronounced differences in the distribution in the lower tail. The relationship between distribution of the bias in schools' reading value-added and the percent of FRPL students in the school is much stronger for reading. Since reading is a skill more strongly related to students' out-of-school experiences (e.g., parents reading to their children), it makes sense that schools' reading value-added is more sensitive to the summer period that their math value-added.

The results from the first research question show that not only are schools' value-add biased by students' summer learning, but also that the practical implications of the bias differentially affect schools based on the types of students they serve. We next evaluate whether the switch from spring-to-spring testing to fall-to-spring testing mitigates this relationship between school demographics and value-add by removing the summer period from the model.

Q2: How correlated are the schools' ACSMs to each other and school demographics?

The descriptive statistics for the SGPs and VAMs are presented in Table 3. The descriptive statistics in Table 3 show that the unconditional distribution of schools' SGPs and value-add are similar across test-timings. Schools' SGPs are centered on 0.5 with a standard deviation of 0.08 for Math and .06 for Reading. The non-parametric nature of the SGPs makes it difficult to interpret the spread of schools' scores. Schools' value-add are mean-zero, with a standard deviation of 1.5 for math and 1.3 for reading. These are approximately 0.1 standard deviations on the spring MAP math and reading assessments indicating that schools have a non-trivial impact on students' math and reading. A one-standard deviation increase in school quality, as measured by value-add, is equivalent to moving a student from the 50th percentile to the 54th percentile in math or reading achievement.

To answer our second question, we explore how closely related the rankings from the four SGPs and two VAMs are within and between test timelines. The correlations are presented in Table 4. We begin with a discussion of the results in Table 4 that pertain to using different models (i.e., SGP vs. VAM) *within* a given test timeline (e.g., spring-to-spring). We do so to contribute to existing work on the sensitivity of school performance measure estimates to model specification. We find that the SGP, the SGP including prior off-subjects, and VAMs generate similar rankings for math with correlations ranging from 0.90 to 0.99. The correlations for reading, however, are much lower across models within a test timeline ranging from 0.6 to 0.9. If reading achievement is more strongly influenced by between student and school factors, then schools' rankings will be more sensitive to model specification.

More germane to the central concern of this paper, we turn to the correlations within the *same* model using different test timelines. 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' rankings within a model between test timelines ranges from 0.95 for Reading SGPs to 0.754 for Math VAMs. Although our school-level correlations are stronger than the year-to-year correlations of teacher value-added, or teachers' within model across timeline correlations (McCaffrey, Sass, Lockwood, & Mihaly, 2009), it is still possible for the discordance across models to be related to student demographics.

The results in Table 5 show the correlations among the ACSMs and aggregate student demographics. We use schools' average achievement levels in 2009-10 as a baseline. Consistent with the extant accountability literature, schools' current math and reading achievement is strongly negatively correlated with the share of FRPL and Minority students in a school ($r \approx -0.57$), and strongly positively correlated with schools' prior spring achievement ($r \approx 0.85$). Using this as a baseline, we show that spring-to-spring SGPs and VAMs are an improvement over proficiency or achievement level ACSMs, especially for math. However, the correlations among the fall-to-spring models and student demographics and prior achievement are all lower than the spring-to-spring models. The fall-to-spring math SGP which also includes students' reading histories has correlations with the share of FRPL and minorities students in a school are -0.1 and -0.2, respectively. If a fall-to-spring test timeline is not possible, states would do well including both subjects as controls in the calculation of schools' SGPs.

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

Interestingly the patterns are a bit mixed for reading. Although most states use only lags of the same subject to generate school SGPs, the correlation between reading SGPs and school demographics that do not include math as a control are approximately -0.4 but it is only approximately -0.25 when math is included as a control. However, the SGPs do not appear to benefit greatly from the inclusion of fall test scores. For the reading only SGP, the correlations drop from -0.43 to -0.34 for the percent of FRPL students in a school and -0.4 to -0.36 for the percent of minority students in a school. The value-added model does benefit from the switch in test timelines. The correlation for the percent of FRPL students in a school changes from -0.44 to -0.23 and for the percent of minority students in a school changes from -0.23 to -0.13. Of the reading models, and similar to the math models, the fall-to-spring VAM is the least related to student demographics.

Q3**: How does schools' relative ranking change between a spring-to-spring and fall-to-spring test timeline?**

Extant research has documented that teachers' relative position within distribution of quality, as measured by VAM, is sensitive to the test timeline (Gershenson & Hayes, 2013; Papay, 2011). However, the research does not differentiate whether the change in the test timeline captures random noise inherent with the estimation of common ACSMs or whether the sensitivity is revealing a source of omitted variable bias. Our last question builds on the prior work by using the kernel density of schools' SGPs and value-add, as well as using conditional transition matrices to evaluate the differential impact of summer learning loss on schools' ACSMs.

To answer our last question, we first plot the kernel density of schools' math and reading value-add and SGPs by tertiles of schools percent of FRPL students, shown in Figure 2 and Figure 3. Figures 2A and 2C show the kernel density of schools' math and reading value-add using a spring-to-spring test timeline, respectively, by the tertiles of the share of FRPL students. Although both subjects show a negative relationship between schools' share of FRPL students and the distribution of value-add, the relationship is more pronounced for reading than for math. The median value-add for the schools in the top tertile of FRPL (e.g., serving the largest shares of FRPL students) is equivalent to approximately the 10^{th} percentile in the value-add distribution for the bottom tertile of FRPL schools. Figures 2B and 2D show the kernel densities for math and reading value-add, respectively, for the fall-to-spring test timeline. Removing the summer from the estimation of schools' math value-add effectively equalizes the distribution of value-add for schools in the middle and top FRPL tertiles. For reading, the median value-add for the top FRPL tertile is now equivalent

to approximately the 30th percentile in the bottom FRPL tertile distribution, a change of 20 percentage points from the spring-to-spring comparison. The results in Figure 3 show similar patterns for the SGP models, although switching to a fall-to-spring test timeline does less to ameliorate the relationship between schools' SGPs and their share of FRPL students.

To further explore the implications of test timelines on schools' accountability standing, we group schools into quintiles based the sample distribution of the percent of FRPL students, a measure of schools' average socio-economic status. We are specifically interested whether the change in schools' relative location in a ACSMs distribution from changing the test timeline is related to school demographics, specifically the percent of FRPL students in the school.

In Table 6 and Table 7 we present the conditional distributions of the FRPL quintiles and the ACSM quintiles for math and reading respectively[vii]. These tables provide a more nuanced look both across models and test timelines of the relationship between ACSMs and school demographics. If a given model was completely independent of schools' share of FRPL students, then we would expect each cell to equal 0.2. Or, said differently, within a given quintile of FRPL we would expect an equal distribution of schools across the quintiles of the ACSM.

As expected from the correlations in Table 5, the results in Table 6 show that schools in the lower quintiles of FRPL are less likely to be in the bottom of the performance distribution and schools in the higher quintiles of FRPL are less likely to be in the top of the performance distribution. For example, only 17, 22, and 21 percent of the schools with the least amount of FRPL students are in the bottom two quintiles of math SGP, SGP off-subject, and VA compared to 37, 46, and 51 percent for schools with the most FRPL students. However, the difference between the two distributions narrows when fall data is incorporated into the ACSMs: 26, 27, and 34 percent of the schools serving the least amount of FRPL students are in the bottom two quintiles of the math SGP, SGP off-subject, and VA, and 42, 41, and 42 percent of the schools serving the most FRPL students are in the bottom quintiles of math SGP, SGP off-subject, and VA. Schools serving the most FRPL students are 11, 11, and 22 percent more likely to be in the bottom two quintiles of math SGP, SGP off-subject, and VA using a spring-to-spring timeline than a fall-to-spring timeline. This is particularly important as accountability systems move from holding all schools accountable for hitting a certain threshold (e.g., 100% proficient) to holding a set number of schools accountable (e.g., the bottom 10%) (Polikoff et al, 2014).

Table 7 shows the same analysis but for reading. A few interesting patterns emerge. The first is the strong relationship between schools' ACSM and FRPL quintiles. For example, 61, 55, and 61 percent of schools serving the most FRPL students are in the bottom two quintiles of the SGP, SGP off-subject, and VA distributions. The numbers drop to 56 , 53 , and 51 percent when incorporating fall achievement data, an improvement of 9 , 9 , and 18 percent. The second is difference between the relationship between the share of FRPL students in a school and the likelihood of being in the top quintile of the ACSM distribution. The results show that 45, 32, and 41 percent of the schools in the top quintile of spring to spring SGP, SGP off-subject, and VA educate the least number of FRPL students, compared to only 11, 19, and 12 percent for schools serving the most FRPL students. The numbers change to 38, 22, and 25 percent for schools serving the least number of FRPL students and 16, 21, and 20 percent for the schools serving the most FRPL students using the fall-to-spring test timeline.

We next compare the percent of schools that either move to a higher quintile, stay within the same quintile, or move to a lower quintile of the ACSM distribution by FRPL quintiles. The results are presented in Table 8. We would expect that the average discordance in spring-to-spring and fall-to-spring ACSM quintiles for schools serving fewer FRPL students would move in a negative direction, and vice versa. For example, if wealthier students gain ground over the summer, then schools serving larger shares of these students will have an upwardly biased spring-to-spring value-add. The results in Table 8 show this exact pattern. Approximately 37, 34, and 38 percent of the schools with the least number for FRPL students are more likely to be in a higher math spring-to-spring SGP, SGP off-subject, and VA quintile than a fall-to-spring quintile, compared to 11, 13, and 20 percent for schools serving the most FRPL students. The opposite pattern occurs when we look at the look at the share of schools that are in a higher fall-to-spring quintile than a spring-to-spring quintile. Nine percent, 14, and 18 percent of the schools serving the least amount of FRPL students are in a higher math fall-to-spring SGP, SGP off-subject, and VA quintile, compared to 31, 28, and 39 percent for schools serving the most FRPL students. The patterns are nearly identical for reading.

The last analysis looks at the misidentification of two groups of schools: 1) schools that are in the bottom of an ACSM distribution using the spring-to-spring test timeline, but are not when the fall-to-spring test timeline is used; and 2) schools that are in the top of the ACSM distribution using the fall-to-spring test timeline, but are not when the spring-to-spring test timeline is used. The former evaluates whether certain types of schools are more likely to be wrongly identified as low-

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

performing under the ACSM. The latter evaluates whether certain schools that should be labeled as high performing are not under the traditional spring-to-spring test timeline. We conduct these analyses separately by FRPL quintile. As indicated by the results in Table 9, the schools in the top FRPL quintile are more likely to be wrongly identified as low-performing under the spring-to-spring test timeline than the other FRPL quintiles, especially the bottom quintile. For example, if states labeled schools in the bottom 20 percent of the math or reading VAM distribution as "failing", 52 and 33 percent of top FRPL quintile schools initially labeled as failing under the spring-to-spring test timeline would be misidentified in math and reading, respectively, compared to just 18 and 11 percent of schools in the bottom FRPL quintile. Similar patterns emerge for the top quintile of the math and reading value-add distribution. Of the schools identified in the top 20 percent of math and reading value-add using the fall-to-spring test timeline, 43 and 57 percent of the top FRPL quintile schools would not be in this category using a spring-to-spring test timeline, compared to 33 and 21 percent for bottom FPRL quintile schools. In sum, the results of our third research question show that students' summer learning experiences induce a bias in ACSMs that introduces inequities into school accountability policies.

Limitations

The results of this paper raise important equity questions related to the decision of school accountability policies. However, there are a number of constraints that should be kept in mind when interpreting the results of this paper. On the one hand, it is important to acknowledge that 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. On the other hand, the Southern state has continued to hire NWEA to administer the MAP assessment to its students in grades 2 through 8 over the past decade. It is likely that at least some schools value the information provided by this assessment, and 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. Lastly, not only does the NWEA data have the appropriate statistical properties for this type of analysis, it is the only data set to our knowledge that includes fall and spring tests for an entire state over a multi-year period and across many grade-levels.

Second, our analysis relies on data from only one state. The demographics in our sample closely resemble many other Southern states. We cannot rule out, however, that the results of this analysis are driven solely by the idiosyncratic educational experiences of the students in our Southern state. It is also possible that the results are driven by unique interactions between the state's standards and the MAP assessment. The NWEA align the questions in the MAP assessment 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 analysis is limited by the simple fact we do not have any indicators for student-level poverty to use in our analysis. 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. It is possible that the inclusion of student-level poverty measures would mitigate the benefit of a switching to a fall-to-spring test timeline.

Lastly, the schools in our analysis did not actually operate under an accountability system that used a SGP or VAM during the school-years used in our analysis. There are potentially important differences between studying the properties of ACSMs 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 SGPs or VAMs would alter the results of this paper. Nonetheless, it is unlikely that schools operating under an accountability system that used growth-based ACSMs would be able to overcome the bias due to summer periods. In fact, high-stakes test generally exacerbate achievement differences between minority/white and low-income/wealthy students, making the results of our analysis a potential lower bound on the true problem.

Discussion and Conclusions

In an effort to improve upon the shortcomings of using student achievement *levels* to hold schools accountable, the federal government and states have incorporated ACSMs that hold schools accountable for student achievement *growth,* usually determined from the amount of learning between spring assessment periods. The spring-to-spring test timeline includes the three month summer period that students spend away from school. Our paper is the first to evaluate the potential

bias from students' differential summer learning introduced into spring-to-spring growth-based ACSMs, and the results have important policy implications.

The first is that while there have been improvements in the design of federal, state, and district school accountability policies, the efficacy of the redesigned federal, state, and district policies is 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 ACSMs. 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 fall-to-spring test timeline doubles the number of test scores available in students' achievement histories for the SGP model. It allows the researcher to essentially pair students with similar fall *and* spring test scores, instead of relying on just prior spring scores. 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.

The second is that, although not the main focus of the paper, our results build on the literature that evaluates the potential of various ACSMs, in this case SGPs and VAMs, to estimate unbiased school effects (Castellano & Ho, 2013; Elhert et al, 2013a; Guarino, Reckase, Stacy, & Wooldridge, 2014). The SGP approach does not directly account for student and school differences in the production of student achievement. It assumes that conditional on students' prior achievement histories, schools are equally able to raise the current achievement level of *any* student regardless of her background. The VAM approach explicitly accounts for student and school differences through the additive use of vectors of student and school demographic control variables. The VAM approach assumes that conditional on students' prior achievement and other differences that are outside of schools' control, schools are equally able to raise the current achievement level of their students. Although the SGP assumptions are often more politically palatable, in no case do they have a weaker correlation between student demographics and schools' rankings than the VAM approach. The continued use of SGPs to hold schools accountable for student achievement, without adjusting for demographic differences among students and schools, will unfairly punish schools educating traditionally underserved students.

Third, states and districts should use as much data as available when estimating ACSMs in a high-stakes context. For example, the modal implementation of SGPs estimates a median SGP for a

given school in the current year using as many years of prior achievement available and does not include the off-subject as a control variable. The method used in this paper pool data over a number of years to estimate our school-level ACSMs. The pooling of data over a number years potentially smooth out year to year variations in student populations, including students' summer experiences. Furthermore, our spring-to-spring SGP model that included the prior achievement histories for both the main and off-subject greatly reduced the correlation between school characteristics and schools' performance rankings. This change can be easily implemented for any accountability system as states are already estimating math and reading school SGPs.

The fourth is that even when the summer period is removed from our SGPs and 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. Elhert et al (2013a) propose using a two-step approach that effectively removes any correlation among school characteristics and schools' math and reading value-add to mitigate the a priori assumption that a teacher or principal is more likely to be labeled as failing at certain types of schools. The method ensures a proportional relationship between school characteristics and schools' ranking within the value-add distribution, but does so potentially at the cost of causality. The results in our paper, and those proposed by Elhert et al (2013a), speak to the need for policy-makers to first figure out what they want to measure—e.g., school's causal effect on students' achievement or a proportional ranking system.

The goal of this paper was threefold. First, we estimated the potential for summer setback to bias commonly used ACSMs. We found that even conditional on prior achievement, student and school characteristics, and school fixed-effects, there is non-zero variation in school-level aggregates of summer learning, and students' summer learning is predictive of their current spring achievement. Second, we demonstrated that ACSMs estimated using a spring-to-spring test timeline are more strongly correlated with schools' demographic characteristics. 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. The results of our paper have important policy implications and raise a number of important design and equity questions related to school accountability policies.

Tables and Figures

Table 1: Student and School Demographics for the 2009-10 School-year

	2009-10 School-year		
Student Demographics	Mean	SD	N
Spring Math MAP Achievement	221.02	19.27	259827
Spring Reading MAP Achievement	212.42	16.75	259327
Fall Math MAP Achievement	213.00	19.92	261638
Fall Reading MAP Achievement	206.56	18.46	261614
Lagged Spring Math MAP Achievement	213.08	20.20	246741
Lagged Spring Reading MAP Achievement	205.97	18.26	246174
White Student	0.528		275997
Black Student	0.362		275997
Hispanic Student	0.061		275997
Mobile Student (Between School-years)	0.086		275997
Mobile Student (Within School-years)	0.028		275997
3rd Grade	0.173		275185
4th Grade	0.176		275185
5th Grade	0.174		275185
6th Grade	0.162		275185
7th Grade	0.161		275185
8th Grade	0.155		275185
% Hispanic in School	0.064		275734
% Black in School	0.364		275734
% White in School	0.530		275734
% FRPL in School	0.580		275734
Urban School	0.165		275997
Suburban School	0.260		275997
Town School	0.154		275997
Rural School	0.421		275997
School Enrollment	671.27	253.03	275997
School Demographics			
% Hispanic in School	0.064		763
% Black in School	0.405		763
% White in School	0.492		763
% FRPL in School	0.626		763
Urban School	0.172		763
Suburban School	0.214		763
Town School	0.157		763
Rural School	0.457		763
School Enrollment	561.50	233.05	763

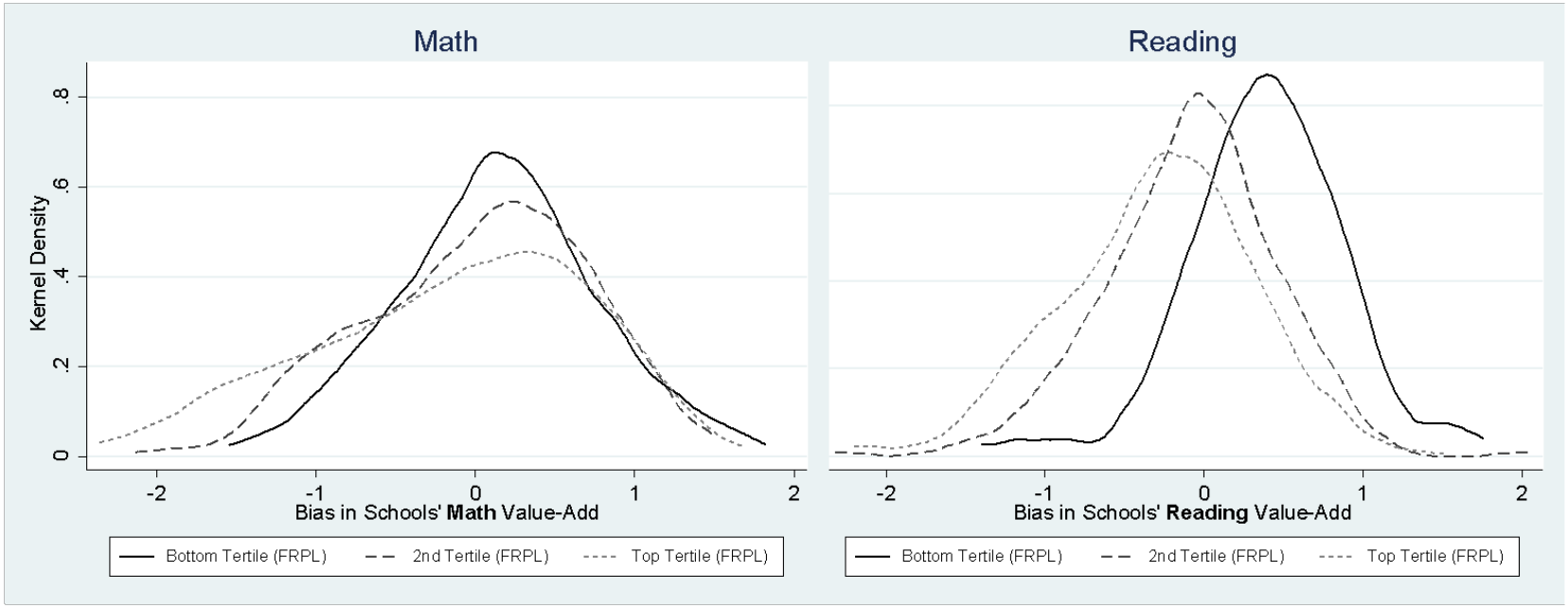
EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

Table 2: Estimates of the Bias in Schools' Math and Reading Value-add due to Summer Setback

	Math	Reading
$P-value: \frac{Cov(\ddot{\delta}_1, Y_{igst}^{Summer})}{Var(\ddot{\delta}_1)} = \frac{Cov(\ddot{\delta}_2, Y_{igst}^{Summer})}{Var(\ddot{\delta}_2)} = \cdots = \frac{Cov(\ddot{\delta}_s, Y_{igst}^{Summer})}{Var(\ddot{\delta}_s)} = 0$	0.000	0.000
$\hat{\theta}_3$	0.520***	0.413***
	(0.002)	(0.002)
Mean$(\hat{\delta}_s - \delta)$	0.009	0.007
SD$(\hat{\delta}_s - \delta)$	0.735	0.617
Corr$\left((\hat{\delta}_s - \delta), \%FRPL\right)$	-0.176	-0.500
Corr$\left((\hat{\delta}_s - \delta), \%Minority\right)$	-0.202	-0.184

Note: The first row is the p-value for the joint F-test that all of the summer setback school fixed-effects are equal to zero. The second row is the estimate coefficient of summer setback, ΔY_{igst}^{Sum}, included in the spring-to-spring VAM specification in Model (2). The third and fourth rows are the mean and standard deviation of the bias in schools math and reading value-add. The fifth and sixth rows are the correlation between schools' math and reading value-add and the percent of FRPL and Minority students in a school, respectively.

Figure 1: Bias in Schools' Math and Reading Value-add due to Summer Learning Loss by Schools' Percent FRPL Tertiles



EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

Table 3: Means and Standard Deviations of Schools' Math and Reading SGPs and Value-added

	Math		Reading	
	Mean	SD	Mean	SD
Student Growth Percentiles (Spring Only)	0.492	0.081	0.490	0.066
Student Growth Percentiles (Fall & Spring)	0.494	0.083	0.491	0.067
Student Growth Percentiles (Spring Only, including off-subject)	0.493	0.082	0.495	0.063
Student Growth Percentiles (Fall & Spring, including off-subject)	0.494	0.083	0.495	0.064
Value-Added Model (Spring to Spring)	0.017	1.463	0.004	1.344
Value-Added Model (Fall to Spring)	0.001	1.438	-0.001	1.261

Table 4: The Correlation among Schools' Math and Reading SGPs and Value-add

		Math (below diagonal) \ **Reading** (above diagonal)					
		Student Growth Percentiles		Student Growth Percentiles (including off-subject)		Value-added Model	
		Spring to Spring	Fall to Spring	Spring to Spring	Fall to Spring	Spring to Spring	Fall to Spring
Student Growth Percentiles	Spring to Spring		**0.947**	0.896	0.836	0.67	0.471
	Fall to Spring	**0.897**		0.875	0.898	0.64	0.508
Student Growth Percentiles (including off-subject)	Spring to Spring	0.978	0.904		**0.918**	0.697	0.506
	Fall to Spring	0.887	0.993	**0.903**		0.637	0.595
Value-added Model	Spring to Spring	0.902	0.844	0.916	0.846		**0.822**
	Fall to Spring	0.720	0.825	0.745	0.835	**0.754**	

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

Table 5: The Correlation Among Math and Reading ACSMs and School Demographics

	Math						
		Student Growth Percentiles		**Student Growth Percentiles (including off subject)**		**Value-Added Model**	
	Current Spring Achievement	Spring to Spring	Fall to Spring	Spring to Spring	Fall to Spring	Spring to Spring	Fall to Spring
% FRPL	-0.575	-0.273	-0.131	-0.197	-0.108	-0.228	-0.125
% Minority	-0.556	-0.281	-0.206	-0.224	-0.192	-0.314	-0.155
Baseline School Achievement	0.867	0.359	0.237	0.317	0.223	0.376	0.262
	Reading						
		Student Growth Percentiles		**Student Growth Percentiles (including off subject)**		**Value-Added Model**	
	Current Spring Achievement	Spring to Spring	Fall to Spring	Spring to Spring	Fall to Spring	Spring to Spring	Fall to Spring
% FRPL	-0.628	-0.427	-0.338	-0.256	-0.183	-0.443	-0.226
% Minority	-0.576	-0.398	-0.359	-0.225	-0.232	-0.229	-0.126
Baseline School Achievement	0.854	0.399	0.360	0.221	0.207	0.305	0.199

Figure 2: Kernel Density Plots of Schools' Spring-to-Spring Math and Reading Value-add

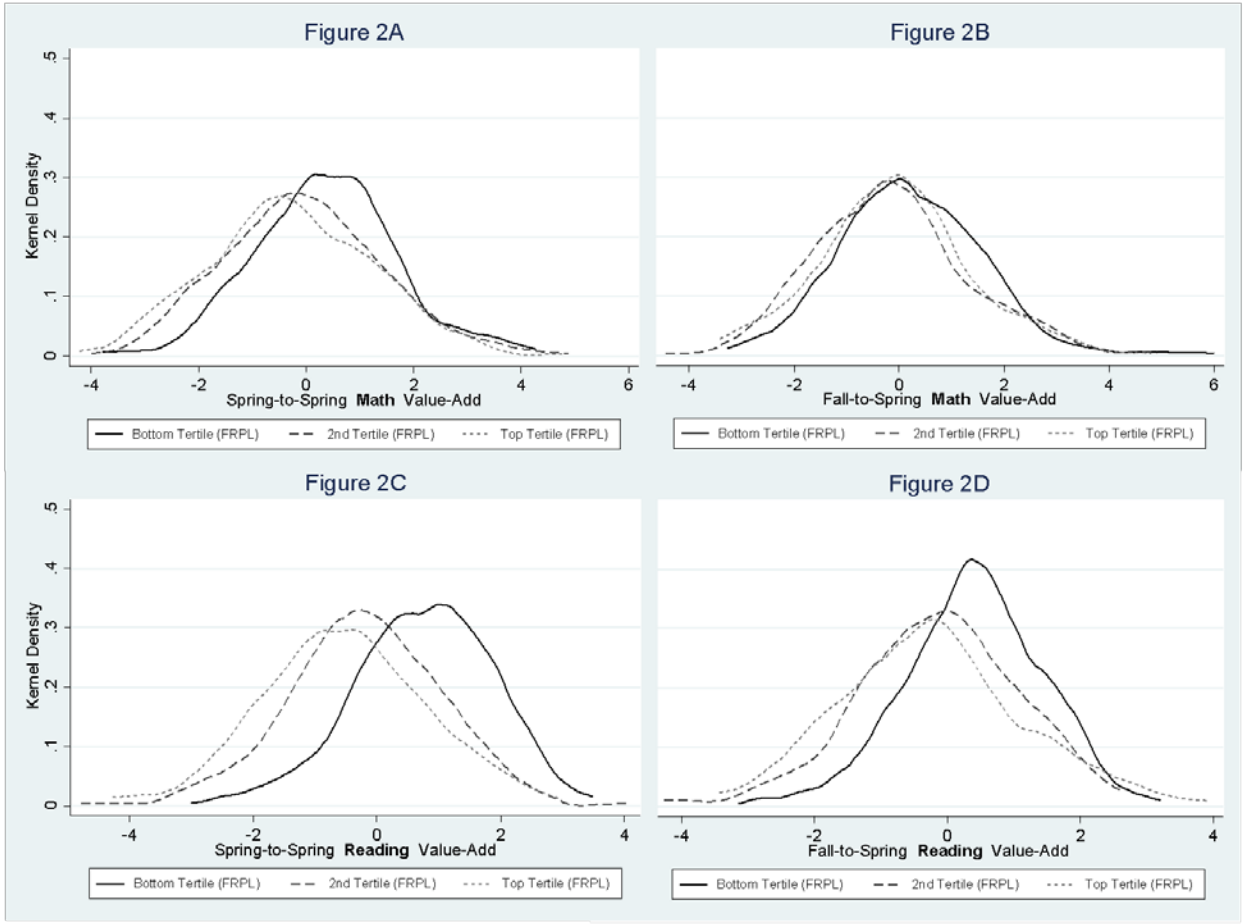
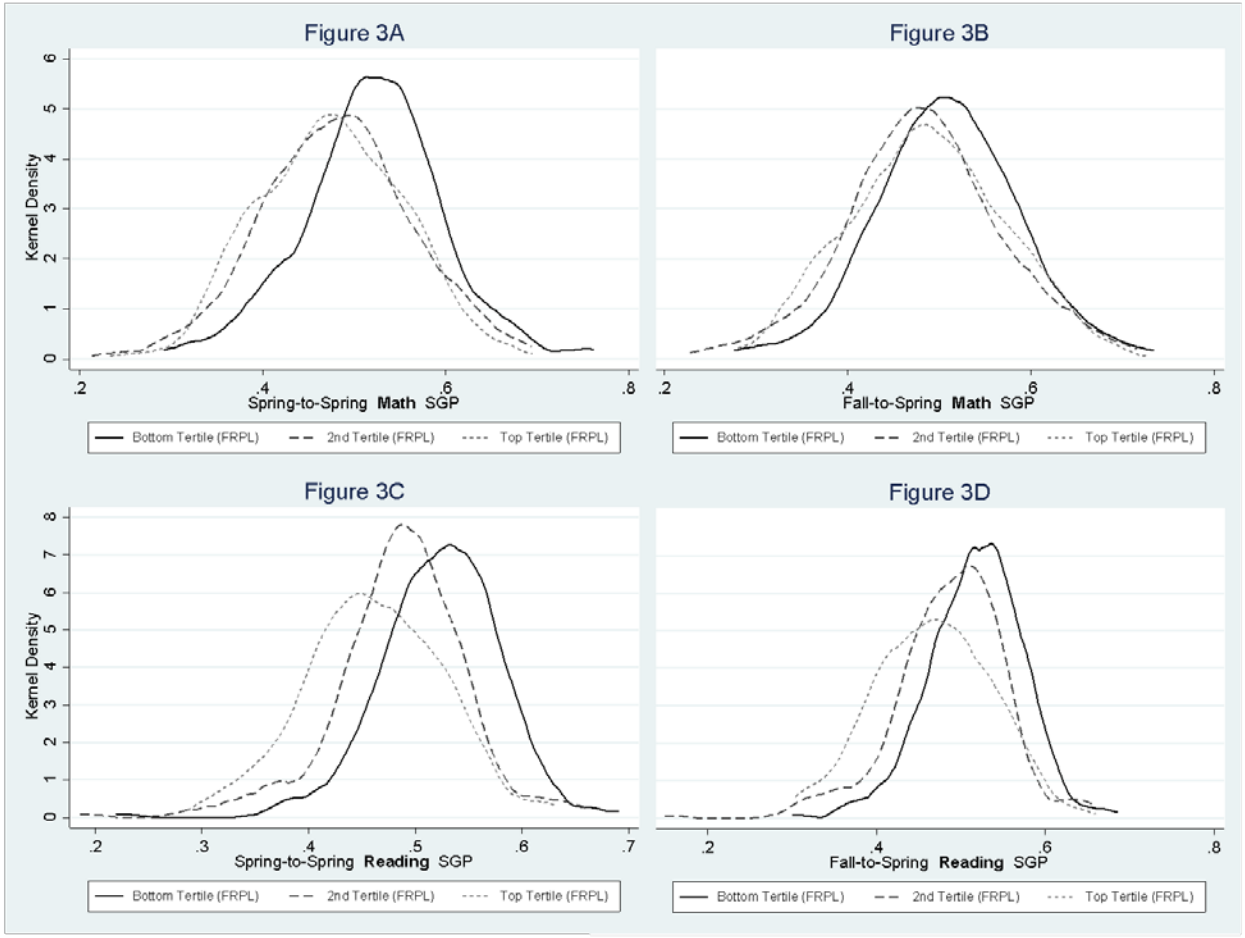


Figure 3: Kernel Density Plots of Schools' Spring-to-Spring Math and Reading SGPs



EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

Table 6: A Transition Matrix of Schools' Math ACSMs by FRPL Quintiles

	Math									
	Spring to Spring					Fall to Spring				
	Bottom Quintile	2nd Quintile	3rd Quintile	4th Quintile	Top Quintile	Bottom Quintile	2nd Quintile	3rd Quintile	4th Quintile	Top Quintile
	Student Growth Percentiles									
Bottom Quintile (% FRPL)	0.065	0.098	0.243	0.296	0.303	0.098	0.157	0.257	0.281	0.212
2nd Quintile (% FRPL)	0.170	0.176	0.191	0.217	0.243	0.170	0.216	0.151	0.203	0.258
3rd Quintile (% FRPL)	0.229	0.261	0.184	0.125	0.204	0.235	0.229	0.197	0.137	0.205
4th Quintile (% FRPL)	0.255	0.275	0.191	0.171	0.105	0.255	0.216	0.237	0.190	0.099
Top Quintile (% FRPL)	0.281	0.190	0.191	0.191	0.145	0.242	0.183	0.158	0.190	0.225
	Student Growth Percentiles (including off-subject)									
Bottom Quintile (% FRPL)	0.085	0.137	0.230	0.296	0.257	0.111	0.157	0.303	0.229	0.205
2nd Quintile (% FRPL)	0.183	0.183	0.178	0.243	0.211	0.190	0.183	0.164	0.216	0.245
3rd Quintile (% FRPL)	0.242	0.229	0.184	0.145	0.204	0.222	0.255	0.171	0.163	0.192
4th Quintile (% FRPL)	0.235	0.248	0.217	0.151	0.145	0.242	0.229	0.211	0.203	0.113
Top Quintile (% FRPL)	0.255	0.203	0.191	0.164	0.184	0.235	0.176	0.151	0.190	0.245
	Value-Added Model									
Bottom Quintile (% FRPL)	0.072	0.144	0.243	0.276	0.270	0.150	0.190	0.164	0.243	0.257
2nd Quintile (% FRPL)	0.203	0.150	0.184	0.257	0.204	0.150	0.190	0.230	0.217	0.211
3rd Quintile (% FRPL)	0.209	0.235	0.197	0.158	0.204	0.255	0.216	0.197	0.112	0.224
4th Quintile (% FRPL)	0.242	0.229	0.211	0.158	0.158	0.235	0.190	0.230	0.237	0.105
Top Quintile (% FRPL)	0.275	0.242	0.164	0.151	0.164	0.209	0.216	0.178	0.191	0.204

Table 7: A Transition Matrix of Schools' Reading ACSMs by FRPL Quintiles

	Reading									
	Spring to Spring					Fall to Spring				
	Student Growth Percentiles									
	Bottom Quintile	2nd Quintile	3rd Quintile	4th Quintile	Top Quintile	Bottom Quintile	2nd Quintile	3rd Quintile	4th Quintile	Top Quintile
Bottom Quintile (% FRPL)	0.046	0.072	0.191	0.243	0.454	0.059	0.124	0.191	0.257	0.375
2nd Quintile (% FRPL)	0.072	0.203	0.224	0.276	0.224	0.085	0.196	0.211	0.283	0.224
3rd Quintile (% FRPL)	0.170	0.235	0.243	0.204	0.151	0.157	0.242	0.243	0.211	0.151
4th Quintile (% FRPL)	0.301	0.288	0.211	0.132	0.066	0.301	0.275	0.197	0.138	0.086
Top Quintile (% FRPL)	0.412	0.203	0.132	0.145	0.105	0.399	0.163	0.158	0.112	0.164
	Student Growth Percentiles (including off-subject)									
Bottom Quintile (% FRPL)	0.059	0.144	0.212	0.268	0.322	0.092	0.163	0.224	0.309	0.217
2nd Quintile (% FRPL)	0.124	0.209	0.212	0.229	0.224	0.111	0.203	0.204	0.217	0.263
3rd Quintile (% FRPL)	0.183	0.196	0.245	0.222	0.158	0.170	0.209	0.283	0.151	0.191
4th Quintile (% FRPL)	0.307	0.216	0.225	0.144	0.105	0.301	0.235	0.145	0.197	0.118
Top Quintile (% FRPL)	0.327	0.235	0.106	0.137	0.191	0.327	0.190	0.145	0.125	0.211
	Value-Added Model									
Bottom Quintile (% FRPL)	0.033	0.072	0.197	0.288	0.414	0.059	0.209	0.192	0.294	0.250
2nd Quintile (% FRPL)	0.098	0.164	0.250	0.248	0.237	0.137	0.137	0.238	0.255	0.230
3rd Quintile (% FRPL)	0.163	0.257	0.224	0.196	0.164	0.176	0.209	0.219	0.209	0.191
4th Quintile (% FRPL)	0.333	0.270	0.171	0.157	0.066	0.346	0.209	0.179	0.131	0.132
Top Quintile (% FRPL)	0.373	0.237	0.158	0.111	0.118	0.281	0.235	0.172	0.111	0.197

Table 8: The Effect of Switching Test Timelines on Schools' Location in the Math and Reading ACSM Distribution

	Math					**Reading**				
	Bottom Quintile of % FRPL Students	Q2	Q3	Q4	Top Quintile of % FRPL Students	Bottom Quintile of % FRPL Students	Q2	Q3	Q4	Top Quintile of % FRPL Students
	Student Growth Percentiles									
Quintile (Spring to Spring) > Quintile(Fall to Spring)	33.3%	22.4%	15.0%	17.8%	11.2%	26.8%	18.4%	15.0%	12.5%	9.2%
Quintile (Spring to Spring) = Quintile(Fall to Spring)	58.2%	61.2%	65.4%	57.2%	57.9%	68.0%	67.8%	68.6%	69.7%	69.1%
Quintile (Spring to Spring) < Quintile(Fall to Spring)	8.5%	16.4%	19.6%	25.0%	30.9%	5.2%	13.8%	16.3%	17.8%	21.7%
	Student Growth Percentiles (including off-subject)									
Quintile (Spring to Spring) > Quintile(Fall to Spring)	34.6%	18.4%	20.3%	19.7%	13.2%	34.6%	18.4%	19.0%	15.1%	15.8%
Quintile (Spring to Spring) = Quintile(Fall to Spring)	51.6%	61.8%	56.9%	63.8%	58.6%	54.2%	57.2%	62.1%	64.5%	62.5%
Quintile (Spring to Spring) < Quintile(Fall to Spring)	13.7%	19.7%	22.9%	16.4%	28.3%	11.1%	24.3%	19.0%	20.4%	21.7%
	Value-added Model									
Quintile (Spring to Spring) > Quintile(Fall to Spring)	37.9%	30.3%	30.7%	30.3%	20.4%	46.4%	18.4%	24.8%	19.7%	11.2%
Quintile (Spring to Spring) = Quintile(Fall to Spring)	43.8%	40.1%	46.4%	40.8%	40.8%	49.0%	67.8%	45.8%	50.7%	52.6%
Quintile (Spring to Spring) < Quintile(Fall to Spring)	18.3%	29.6%	22.9%	28.9%	38.8%	4.6%	13.8%	29.4%	29.6%	36.2%

Table 9: The Share of Schools Consistently Identified as Low- or High-Performing by FRPL Quintile

Math	Bottom FRPL Quintile	2nd Quintile	Third Quintile	Fourth Quintile	Top FRPL Quintile
Bottom Quintile of **SGP** (Spring-to-Spring)	0.00	0.20	0.17	0.23	0.26
Top Quintile of **SGP** (Fall-to-Spring)	0.09	0.21	0.23	0.27	0.47
Bottom Quintile of **SGP off-subject** (Spring-to-Spring)	0.23	0.25	0.25	0.14	0.21
Top Quintile of **SGP off-subject** (Fall-to-Spring)	0.13	0.30	0.17	0.12	0.41
Bottom Quintile of **Value-add** (Spring-to-Spring)	0.18	0.53	0.43	0.43	0.52
Top Quintile of **Value-add** (Fall-to-Spring)	0.33	0.34	0.34	0.20	0.43
Reading	Bottom FRPL Quintile	2nd Quintile	Third Quintile	Fourth Quintile	Top FRPL Quintile
Bottom Quintile of **SGP** (Spring-to-Spring)	0.00	0.27	0.19	0.17	0.10
Top Quintile of **SGP** (Fall-to-Spring)	0.09	0.24	0.26	0.31	0.48
Bottom Quintile of **SGP off-subject** (Spring-to-Spring)	0.11	0.21	0.21	0.15	0.18
Top Quintile of **SGP off-subject** (Fall-to-Spring)	0.21	0.33	0.28	0.33	0.25
Bottom Quintile of **Value-add** (Spring-to-Spring)	0.11	0.21	0.21	0.27	0.33
Top Quintile of **Value-add** (Fall-to-Spring)	0.21	0.33	0.28	0.60	0.57

References

Authors, 2014

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.

Baker, G. P. (2000). The Use of Performance Measures in Incentive Contracting. The 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'slow-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.

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

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.

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

Chetty, R., Friedman, S.N., & Rockoff, (2013). Measuring the impacts of teachers I: Evaluating the bias in teacher value-added estimates. NBER Working Paper 19423.

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.

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. (2013a). *Selecting growth measures for school and teacher evaluations*. Department of Economics Working Paper Series, (WP 12-10). University of Missouri.

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. (2014). The implications of summer learning loss for value-added estimates of teacher effectiveness. AEFP 2014.

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., Wooldridge, J. (2012). *Can value-added measures of teacher performance be trusted?* The Education Policy Center at Michigan State University Working Paper (#18). Michigan State University.

Guarino, C., Reckase, M., Stacy, B., Wooldridge, J. (2014). *A comparisonof growth percentile and value-added models of teacher performance*. The Education Policy Center at Michigan State University Working Paper (#39). Michigan State University.

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.

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.

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.

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.

Molfese, V. J., Modglin, A., & Molfese, D. L. (2003). The role of environment in the development of reading skills: A longitudinal study of preschool and school-age measures. *Journal of learning disabilities, 36*(1), 59-67.

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. (2010). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal.*

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

Reinstein, D. A., & Snyder, C. M. (2005). The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. The Journal of Industrial Economics, 53(1), 27-51.

Riddle, W., & Kober, N. (2011). State policy differences greatly impact AYP Numbers (pp. 1-22). Wasington, D.C.: Center on Education Policy.

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.

Scherrer, J. (2011). Measuring Teaching Using Value-Added Modeling: The Imperfect Panacea. *NASSP Bulletin*.

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.

[i] We do not standardize students' Read or Math MAP scores because the scores are on a vertical and interval scale, and are normally distributed.

[ii] In fact, we include a separate "days" variable for every school year t included in the model. These variables are indexed j=0 to 3, and each captures the number of days between the given spring score in year t-j) and the last spring test for each of the school years included in the model, up to school year t-3.

[iii] One reason SGPs have gained traction with states is that that the federal government does not allow the use of student or school demographics in growth models.

[iv] Although all students in the state take the MAP tests in the fall and spring, schools have flexibility on when they administer the exams. In order to hold schools accountable only for the amount of potential learning available to the students, we control for the number of instructional days available to the students in school *s* between the prior focus test score and the current test score. For test timings that span the summer, we subtract 92 days. For example, the number of instructional days for the spring-to-spring timing is the date of the current administration minus the date of the prior spring administration minus 92. Extant research finds that students' within school-year achievement growth is approximately linear (Fitzpatrick, Grissmer, & Hastedt, 2011). The results of our analysis are unchanged if we use a higher-order polynomials in the days between tests.

[v] Note that it could be confusing to keep track of whether each summer period belongs to school-year t or t-1, since summers by definition occur between two school years. In this paper, we

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**

consistently denote the summer prior to school-year t as belonging to school-year t. For instance, let t equal the 2010-11 school year. In this case, the corresponding summer for school-year t, ΔY_{igst}^{Sum}, is the summer just before school-year 2010-11. In contrast, ΔY_{igst-1}^{Sum}, would refer to the summer just before the 2009-10 school year.

[vi] Technically the term $\ddot{\delta}_s$in this expression, is actually the residual of a school indicator variable for school s regressed on all the other independent variables in equation(2). To keep the presentation streamlined we did not differentiate between the indicator variable used as the fixed-effect and its estimate coefficient.

[vii] The cutoffs for the quintiles of FRPL status are (from bottom to top): 31.9%, 52.4%, 64.5% 75.5%, and 88.8%.

EdPolicyWorks Working Paper Series No. 25. May 2014.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | **University of Virginia**