The Implications of Summer Learning Loss for Value-Added Estimates of Teacher Effectiveness

Seth Gershenson, American University and IZA♦
Michael S. Hayes, Rutgers University - Camden

This article is forthcoming in *Educational Policy*. The appropriate citation is:

Gershenson, Seth, & Hayes, Michael S. 2016. The implications of summer learning loss for value-added estimates of teacher effectiveness. In press, *Educational Policy*. DOI: 10.1177/0895904815625288

Abstract

School districts across the United States increasingly use value-added models (VAMs) to evaluate teachers. In practice, VAMs typically rely on lagged test scores from the previous academic year, which necessarily conflate summer with school-year learning and potentially bias estimates of teacher effectiveness. We investigate the practical implications of this problem by comparing estimates from "cross-year" VAMs to those from arguably more valid "within-year" VAMs using fall and spring test scores from the nationally representative Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K). "Cross-year" and "within-year" VAMs frequently yield significant differences that remain even after conditioning on participation in summer activities.

Keywords: Value-added models; teacher quality; summer learning loss

♦ Corresponding author. Email: gershens@american.edu. The authors thank three anonymous referees and conference participants at the 2014 annual meetings of the American Educational Research Association and Association for Education Finance and Policy for providing helpful feedback. Any remaining errors are our own.

Closing persistent achievement gaps between students of different demographic and socioeconomic backgrounds is a primary goal of education policy in the United States. A growing consensus agrees that providing high-quality teachers to all students must play a prominent role in reaching this goal, though identifying effective teachers is difficult in practice (Baker et al., 2010; Harris, 2011; Nye et al., 2004). Value-added models (VAMs) that attempt to identify individual teachers' contributions to gains in student achievement are gaining popularity, but remain controversial (Baker et al., 2010; Chetty, Friedman, & Rockoff, 2014; Glazerman et al., 2010; Harris, 2011; Hill, 2009; Kelly, 2012; McCaffrey et al., 2003; Papay, 2011).

Intuitively, VAMs use previous achievement (lagged test scores) as a sufficient statistic, or proxy, for the unobserved history of family, educational, and individual inputs received by children (Harris, Sass, & Semykina, 2014; Todd & Wolpin, 2003). Doing so is important, as a consensus agrees that teachers should not be held accountable for student characteristics, such as past inputs, that are outside teachers' control (Baker et al., 2010; Harris, 2011). A similar argument applies to controlling for current inputs that are outside teachers' control (e.g., class size). Therefore, to produce unbiased estimates of teacher effects that can be given a causal interpretation, VAMs must control for all "current" inputs received by the child *after* they took the test that proxies for the unobserved historical inputs received by the child (i.e., the VAM's lag score). Failing to adequately control for such inputs will potentially underestimate the effectiveness of teachers who teach relatively disadvantaged students (e.g., students who experience less supportive home and neighborhood environments), and vice versa.

However, most VAM-based analyses rely on standardized tests that are administered once per year, either in the fall or spring (Downey, von Hippel, & Hughes, 2008; Harris, 2009; Papay, 2011; Winters & Cowen, 2013). For example, the value-added components of recent

teacher-assessment programs in Houston and Nashville relied on student assessments administered each spring. As a result, the lag score that proxies for unobserved historical inputs comes from the previous academic year and thus fails to control for inputs received, and learning that occurred, during the summer vacation. Students' exposure to stimulating activities and supportive environments outside of the traditional school day, particularly during summer vacation, are outside of teachers' control and should be controlled for in VAMs (Linn, 2009). For example, differences by socioeconomic status (SES) in children's summer time use and exposure to parental involvement (Gershenson, 2013), participation in enriching summer activities (Chin & Phillips, 2004), and summer learning rates (Alexander, Entwisle, & Olson, 2001; Cooper et al., 1996) are well documented.

Differential rates of summer learning threaten the validity of VAM-based analyses of the relationship between school inputs such as teachers and student achievement, as the common practice of evaluating students' achievement growth from one academic year to the next necessarily conflates students' summer learning with learning that occurred during the school year (Baker et al., 2010; Linn, 2009; Papay, 2011). Downey et al. (2008) and McEachin and Atteberry (2014) raise the same concerns regarding the evaluation of school performance. Specifically, the problem arises because the administrative data used to fit VAMs generally span the summer vacation but rarely, if ever, contain information on students' summer activities. The current study contributes to our understanding of the validity, robustness, and best practices of VAM-based analyses of teacher effectiveness by investigating the practical implications of measuring achievement only once per academic year in the absence of data on summer activities.

The empirical analysis utilizes both fall and spring test scores of the kindergarten and first-grade students surveyed by the nationally representative Early Childhood Longitudinal

Study – Kindergarten Cohort (ECLS-K). Specifically, we examine the validity of VAM-based rankings of classroom effects generated by VAMs that rely on annual test scores but fail to control for summer activities. Intuitively, we compare rankings generated by "cross-year" VAMs that rely on either spring-to-spring or fall-to-fall gains in achievement to analogous rankings generated by arguably more valid "within-year" VAMs that rely on fall-to-spring achievement gains. Because previous research suggests that summer learning rates are correlated with observed student and household characteristics, we also examine the ability of the student characteristics typically observed in administrative data, as well as richer measures of household characteristics and children's summer activities typically unavailable in administrative data, to control for the summer learning inherent in "cross-year" VAMs (e.g., McCaffrey et al., 2003).

Theoretical Background and Literature Review

Summer Learning Loss

Education researchers have long been interested in summer vacation's effect on learning, which has also been referred to as "summer setback," "summer learning loss," and the "summer slide." Several empirical studies have investigated the magnitude and correlates of summer learning loss; see Cooper et al. (1996) and Borman and Boulay (2004) for thorough reviews of this literature. On average, studies conducted prior to the 1970s generally found negative effects of summer vacation on math achievement and either no or mixed effects on reading and literacy achievement (Cooper et al., 1996). However, these early studies failed to account for heterogeneity in summer learning rates, which may have masked differences in the summer learning rates of students from different demographic and socioeconomic backgrounds. Of

course, summer learning rates are likely to vary across students for a variety of reasons (Gershenson, 2013), which may partially explain the mixed results in earlier studies.

Borman et al. (2005) discuss four potentially interrelated mechanisms that may cause children in low-SES households to experience smaller achievement gains during the summer vacation than their more advantaged counterparts. First, investment models hypothesize that high-SES parents have the time and financial resources to invest in the development of children's human capital during the summer vacation (Becker & Tomes, 1986). Investment models are conceptually similar to the "faucet theory" of Entwisle, Alexander, and Olson (2001), which posits that SES differences in summer learning rates are driven by high-SES households being better able to compensate when the flow of resources from the "school tap" is shut off. Second, SES differences in summer learning rates may result from different parenting strategies (Entwisle, Alexander, & Olson, 1997; Heyns, 1978; Lareau, 2003). Third, psychological models hypothesize that high-SES parents have higher expectations for children's achievement and behavior, which may lead to higher rates of summer learning (Entwisle et al., 1997; Hoover-Dempsey & Sandler, 1995). Finally, heterogeneity in either access or returns to participation in organized summer activities may exacerbate differences by SES in summer learning rates (Cooper et al., 2000).

More recent studies of summer learning have documented significant differences by SES in the development of reading and literacy skills during summer vacation (e.g., Burkam et al., 2004; Cooper et al., 1996; Downey, von Hippel, & Broh, 2004; Alexander et al., 2001; Heyns, 1978). That these studies generally find differences by SES in summer reading and literacy gains but not in math gains is consistent with the finding that school inputs have relatively greater effects on math achievement than on reading achievement (e.g., Hanushek & Rivkin, 2010;

Jacob, 2005; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). This may result from children being more likely to develop reading and literacy skills than math skills at home (Currie & Thomas, 2001) and high-SES households spending more time reading to/with children (Phillips, 2011).

However, the R^2 of summer learning regressions than condition on student and household covariates are relatively small, suggesting that observed household characteristics and students' summer activities only explain 8 to 13 percent of the variation in summer learning (e.g., Burkam et al., 2004; Downey et al., 2004). We return to this point below when considering the ability of statistical controls to control for summer learning in VAMs. Specifically, the empirical analysis will investigate the ability of observed student and household characteristics, as well as students' summer activities, to control for the "summer learning bias" inherent in "cross-year" VAMs.

Test Timing and Value-Added Models

Numerous states and school districts now use estimates of teacher effectiveness generated by value-added models (VAMs) to make moderate-stakes decisions regarding teacher retention and merit pay (Armour-Garb, 2009; Harris et al., 2012; Kelly, 2012). Accountability policies often use similar methods to evaluate schools (McEachin & Atteberry, 2014). Most of these VAM-based analyses rely on standardized tests that are administered once per academic year, either in the fall or spring (Downey et al., 2008; Harris, 2009; McEachin & Atteberry, 2014; Papay, 2011; Winters & Cowen, 2013). Critics of these policies have stressed the problems associated with measuring achievement gains between, as opposed to within, academic years (Baker et al., 2010; Linn, 2009; Papay, 2011). Downey et al. (2008) and McEachin and Atteberry (2014) make a similar point regarding the use of VAMs to evaluate school-level performance,

and find nontrivial differences between measures of school performance that do and do not adjust for summer learning. Specifically, whether tests are administered each fall or each spring, the "cross-year" gain necessarily includes the gains and losses experienced during summer vacation (Downey et al., 2008). This is problematic if, as is likely the case, summer learning rates (i.e., students) are not randomly distributed across classrooms (Dieterle, Guarino, Reckase, & Wooldridge, 2015; Papay, 2011; Rothstein, 2010) and data on children's summer activities are not available. For example, parental involvement, which is not observed in most administrative datasets, might create variation across classrooms in average summer learning as parental involvement potentially affects both summer learning (Gershenson, 2013) and classroom assignments (Dieterle et al., 2015). However, even if summer learning gains and losses are randomly distributed across students and classrooms, the mere presence of summer learning in cross-year VAMs adds noise to the error term that increases the probability that VAMs misclassify teachers. For these reasons, whenever variation in summer learning rates is present, within-year VAMs likely yield more accurate measures of teacher effectiveness than do cross-year VAMs.

Figure 1 shows how variation in summer learning can bias cross-year value-added estimates of teacher effectiveness. Suppose that there are two first-grade teachers who are equally effective and that each teacher is assigned one student (or one classroom). At the first time point depicted in figure 1, the spring of kindergarten, the two students (or classrooms) have identical achievement levels (test scores = 3). However, during the summer vacation between kindergarten and first grade, student A continued learning while student B experienced summer learning loss. The dashed lines' slopes represent the students' summer learning rates. As a result, students A and B entered first grade with test scores of 4 and 2, respectively. The solid lines'

slopes represent the students' first-grade (within-year) learning rates. Assuming that the two students have identical propensities for school-year learning and that all other schooling and home inputs are held constant, the solid lines' slopes also represent teacher effectiveness. However, as discussed above, many states and districts use the previous spring's test score as the lag score in VAMs and instead measure teacher effectiveness by taking the slope of the dotted lines (i.e., cross-year VAMs). As figure 1 makes clear, the cross-year VAM incorrectly indicates that student A's teacher is more effective than student B's teacher, as the dotted line for student A is steeper than that for student B. The actual magnitude of the bias might be even larger if, as the result of summer learning loss, B's teacher spends more time re-teaching concepts from the previous academic year. This would limit B's within-year academic growth, at no fault of the teacher, and lead to a flattening of B's solid line in figure 1.

The bias caused by variation across classrooms in summer learning loss could be significantly reduced either by administering tests on the first and last days of the academic year being tested or by controlling for rich measures of children's summer activities and time use in cross-year VAMs. The former is impractical, however, as it would further increase the time and resources devoted to testing and potentially incentivize teachers to game the system by artificially depressing fall scores (Baker et al., 2010). Moreover, even if teachers did not strategically depress fall test scores, teachers who are effective early in the school year would not receive credit for student learning that occurred prior to the fall baseline test, and would actually be harmed by the resulting higher fall baseline scores in value-added analyses of teacher effectiveness. Nor is collecting and controlling for detailed data on students' activities, parental involvement, and time use during the summer vacation a panacea, as doing so would be similarly costly and politically contentious (von Hippel, 2009). Accordingly, in the context of existing

policies and data limitations, the first-order policy-relevant questions regard the validity of estimates of teacher effectiveness generated by "cross-year" VAMs and the ability of basic demographic and SES variables to control for the differences in summer learning and home environments inherent in cross-year VAMs.

To date, Papay (2011) is the only study to have empirically investigated the implications of summer learning for VAM-based rankings of teacher effectiveness. Using six years of matched student-teacher data on third- through fifth-grade student test scores from a large urban district in the Northeastern U.S., Papay estimated Spearman Rank Correlations between teacher-effectiveness rankings generated by fall-to-spring and spring-to-spring gains on Scholastic Reading Inventory (SRI) tests of about 0.7. This suggests that the two rankings are highly, but not perfectly, positively correlated. The point estimate of 0.7 falls at the high range of comparable estimates of the stability of VAM-based rankings across both time and subjects (Goldhaber & Hansen, 2013; Loeb & Candelaria, 2012; Loeb, Kalogrides, & Beteille, 2012; McCaffrey et al., 2009). Critics of VAMs consider these correlations too low to justify using VAM-based estimates of teacher effectiveness to make high-stakes decisions (Hill, 2009).

The current study extends Papay's (2011) analysis of the robustness of VAM-based rankings of teacher effectiveness to differences in test timing in several ways. First, nationally-representative ECLS-K data provide results that are more generalizable to smaller and non-urban districts. Second, we consider richer comparisons of the resulting rankings that provide a more nuanced view of the instability between "within-year" and "cross-year" VAMs, which we describe in the methods section. Third, we formally investigate the ability of conditioning on the student characteristics typically available in administrative datasets, as well as data on household characteristics and students' summer activities, to mitigate the bias attributable to the summer

learning inherent in "cross-year" VAMs. Finally, because collecting detailed summer-activity data and testing at the start and end of the school year are arguably impractical in the immediate future, we consider whether fall-to-fall or spring-to-spring VAMs are potentially more valid.

Data

The current study utilizes data from the nationally representative Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K). The ECLS-K is a longitudinal data set collected by the National Center for Education Statistics (NCES). The original sample of approximately 22,000 children from about 1,000 kindergarten programs was designed to be nationally representative of the cohort that began kindergarten in the 1998-99 academic year. The cohort nature of the ECLS-K means that teachers are only observed in one school year, thus we focus on estimating "classroom" rather than teacher effects, as further discussed in the methodology section. Because the ECLS-K oversampled certain subgroups of the population, all analyses are conducted using NCES-provided sampling weights that adjust for the survey's nonrandom sampling frame. However, as suggested by Solon, Haider, and Wooldridge (2015), un-weighted estimates are considered as part of the sensitivity analysis.

All children in the initial sample were surveyed in the fall and spring of kindergarten and the spring of first grade. However, the analytic sample is restricted to the 30 percent random subsample of children who were also surveyed in fall of first grade [N = 4,150].[1] The fall of first grade observations are crucial to the analyses described in the methodology section, as they facilitate the following calculations and comparisons:

- Test-score change between spring of kindergarten (K) and fall of first grade (1)

- Test-score change between Fall K and Fall 1, versus between Fall K and Spring K
- Test-score change between Spring K and Spring 1, versus between Fall 1 and Spring 1

We further restrict the analytic sample by excluding students who repeated kindergarten or first grade [N = 3,600], changed schools between kindergarten and first grade [3,500], experienced a mid-year classroom change [N = 3,450], were missing basic demographic data or classroom indicators [N = 2,450 first graders; N = 2,650 kindergarteners], or were in a classroom in which fewer than five classmates were sampled by the ECLS-K, which results in final analytic samples of 1,250 first graders and 1,500 kindergarteners. School changers are excluded to avoid conflating the impact of summer learning loss with that of changing schools, though it is worth noting that including school changers in the analytic sample and conditioning on a "changed schools" indicator yields qualitatively similar results. The last restriction increases the precision of estimated classroom effects, which are the parameters of primary interest in the current study, though we relax this assumption in the sensitivity analysis. The baseline analytic samples include 100 unique schools, 150 first-grade classrooms, and 200 kindergarten classrooms.

The ECLS-K data are well suited for an investigation of the practical implications of differential summer learning rates for value-added estimates of classroom effects for three general reasons. First, while data spanning multiple summer vacations would be ideal for reducing sampling error (Koedel & Betts, 2011) and tracking trends in summer learning rates, the ECLS-K is the only nationally representative survey of U.S. students that contains both fall and spring test scores spanning even one summer that also links students to classrooms. Moreover, this enables the estimation of both "within-year" and "cross-year" VAMs using data

on the same students, which ensures that the results are not driven by changes in the composition of teachers' classrooms.

Second, the ECLS-K contains data on students' summer activities (e.g., participation in organized summer activities, summer school attendance, trips to the library, math and reading practice at home), which facilitate tests of the ability of data on summer activities to reduce the bias inherent in "cross-year" VAMs.

Third, the fall and spring tests administered by the ECLS-K covered the same content and were not associated with any stakes or accountability programs, so teachers had no incentive to strategically divert resources or instructional time towards a specific test (Fitzpatrick, Grissmer, & Hastedt, 2011). Specifically, the ECLS-K administered age-appropriate reading and mathematics tests during each wave of the survey that were modelled after other early childhood tests that have previously been used in value-added style analyses of educational interventions (e.g., Peabody Picture Vocabulary Test [PPVT]; Test of Early Math Ability [TEMA]) (Rock & Pollack, 2002). The math examinations tested children's abilities in the following subjects: numbers and shapes, relative size, ordinality and sequence, addition and subtraction, and multiplication and division. The reading examinations tested children on letter recognition, beginning sounds, ending sounds, sight words, and words in context. Because the achievement tests used a two-stage assessment approach, all children did not take identical exams. Hence, the ECLS-K computed vertically scaled test scores based on the full set of test items using Item Response Theory (IRT) (Rock & Pollack, 2002). The ECLS-K assessments are reliable, as evidenced by IRT theta coefficients of internal consistency greater than 0.92 in all waves and subjects (Rock & Pollack, 2002). Similarly, our application of the test-retest method for identifying overall test measurement error proposed by Boyd et al. (2013) yields reliability

estimates 0.87 and 0.95 for spring-K and fall-1 math and reading tests, respectively. In baseline analyses test scores are standardized by subject, grade, and wave (fall or spring) to have means of zero and standard deviations of one using the full ECLS-K sample (Ballou, 2009). However, the main results are robust to measuring student achievement using either unstandardized vertically scaled test scores, or the ECLS-K's theta estimates of latent student ability preferred by Quinn (2015), as shown in a sensitivity analysis.[2]

Finally, an important caveat to the current study is that in both the fall and spring semesters, ECLS-K tests were administered to different students on different days (Fitzpatrick et al., 2011). Differences in test dates across schools, classrooms, and even students within classrooms are common in the data, as a relatively small number of ECLS-K administrators individually met with each student. To avoid conflating summer learning with school-year learning that occurred either before the fall test or after the spring test, we adjust all subsequent analyses by controlling for the number of days prior to the fall test and after the spring test. Fitzpatrick et al.'s (2011) study of the effect of time spent in formal schooling on academic achievement finds no evidence of nonlinear effects of days in school on achievement. Moreover, Fitzpatrick et al. find that ECLS-K test dates are essentially randomly distributed across students, suggesting that the differences in ECLS-K assessment dates do not invalidate the current study.

The extent to which VAM-based estimates of teacher effectiveness are biased by differential rates of summer learning depends upon the distribution of summer learning across students and classrooms. Table 1 describes the summer learning in both math and reading that occurred during the summer between kindergarten and first grade for the children of the ECLS-K. Recall that the test scores were standardized using all available test score data, so the means and standard deviations (SD) are not precisely 0 and 1 in the analytic sample. The summer

learning losses (gains) analyzed in table 1 are adjusted for the exact timing of the test to account for the fact that some kindergartners took the test well in advance of the end of kindergarten and some first graders took the test well after the start of first grade.[3] The average student lost nearly one half of a test-score standard deviation (SD) in both subjects. Moreover, the estimated SD of summer learning are approximately 0.5 SD as well, suggesting that there is considerable variation in summer learning rates across students.

To investigate how summer learning is distributed across schools and classrooms, table 1 also reports within-school and within-classroom SD. Variation in summer learning across schools, classrooms, and students is further summarized by estimating the overall, within-school, and within-classroom SD of summer achievement gains. The "within" SD are estimated by the SD of the residuals from regressions of summer achievement on sets of either school or classroom fixed effects (FE). The within-school SD are about 90 percent as large as the overall SD, indicating that approximately 90 percent of the variation in summer learning exists within as opposed to between schools. This within-school variation could be within or between classrooms (teachers). In the context of estimating classroom effects using annual test-score data, the latter would be particularly troublesome. However, the within-classroom SD are quite close to the within-school SD, indicating that approximately 98 percent of the within-school variation in summer learning exists within, as opposed to between, classrooms.

While means and SD provide useful summaries of the variation in summer learning rates, it is also instructive to consider the entire distribution. Accordingly, histograms of the summer gains and losses in children's math and reading achievement observed between the spring of kindergarten and the fall of first grade are plotted in figures 2.A and 2.B, respectively. The distributions of summer learning in both subjects are approximately symmetrical, centered on the

mean, and contain a nontrivial fraction of students who experienced either substantial gains or losses larger than 0.5 test-score SD. These figures reinforce the idea that nontrivial variation exists in students' summer learning. Similarly, the histograms depicted in figure 3 collapse the student-level data to the classroom level and plot the distributions of classroom mean and median summer gains (losses). For both math and reading, whether measured by the classroom mean or median, the histograms in figure 3 indicate nontrivial variation in summer learning across classrooms, potentially biasing cross-year VAM estimates of classroom effectiveness.

Table 2 summarizes the students who comprise the analytic sample, as well as their summer activities. The analytic sample is about 80% white, 7% black, and 8% Hispanic. About half the students are female and about 10% reported residing in households below the poverty line. The poverty indicator may be an important proxy for children's exposure to enriching items and activities, as Kaushal, Magnuson, & Waldfogel (2011) document an "income gap" in children's participation in organized activities, among other things. Similarly, Gershenson (2013) shows that children in low-income households watch significantly more television and engage in significantly less conversation with adults than their wealthier counterparts during the summer vacation. About 30% of students attended a private school and urban, suburban, and rural schools are approximately equally represented.

In addition to these demographic and school-type variables that are typically available in administrative data, the ECLS-K also contains rich data on mothers' educational attainment and children's summer activities that may be related to summer learning. For example, over 30% of mothers hold a four-year college degree. This is a potentially important control variable, as the literature on parental involvement finds that highly-educated parents spend more time interacting with their children (Guryan, Hurst, & Kearney, 2008) and Gershenson (2013) shows that such

gaps are even larger during the summer vacation. Similarly, nearly two thirds of children in the analytic sample participated in organized summer activities, which previous research has shown to predict academic achievement (Covay & Carbonaro, 2010). Finally, the majority of mothers reported frequently practicing math and reading to/with their children during the summer vacation, the latter of which previous research has also shown to predict children's cognitive development (e.g., Phillips, 2011).

Methodology

The ECLS-K data enable two sets of comparisons. For kindergarteners, we compare VAM-based rankings of classroom effectiveness generated by fall-to-spring (within-year) achievement gains to the potentially less-valid rankings generated by fall-to-fall (cross-year) gains; the corresponding VAM specifications are given by equations (1a) and (1b), respectively:

$$A_{ic}^{Spring,K} = \gamma + \alpha A_i^{Fall,K} + \boldsymbol{\beta}\mathbf{x}_i^K + \theta_c^K + u_{ic} \quad (1a)$$

and

$$A_{ic}^{Fall,1} = \gamma + \alpha A_i^{Fall,K} + \boldsymbol{\beta}\mathbf{x}_i^K + \theta_c^K + u_{ic}. \quad (1b)$$

Similarly, for first graders, we compare VAM-based rankings of classroom effectiveness generated by fall-to-spring achievement gains to the arguably less-valid rankings generated by spring-to-spring gains, as shown in equations (2a) and (2b), respectively:

$$A_{ic}^{Spring,1} = \gamma + \alpha A_i^{Fall,1} + \boldsymbol{\beta}\mathbf{x}_i^1 + \theta_c^1 + u_{ic} \quad (2a)$$

and

$$A_{ic}^{Spring,1} = \gamma + \alpha A_i^{Spring,K} + \boldsymbol{\beta}\mathbf{x}_i^1 + \theta_c^1 + u_{ic}. \quad (2b)$$

In equations (1) and (2), students and classrooms are indexed by *i* and *c*, respectively; *K* and 1 indicate kindergarten and first-grade, respectively; *A* is academic achievement (i.e., standardized math and reading scores); the vector **x** contains some combination of the student characteristics and summer activities described in table 2; θ are the classroom FE upon which rankings of classroom effectiveness will be based; and *u* is a mean-zero error term that captures the unobserved predictors of achievement (e.g., unobserved household shocks, neighborhood effects, illness, parental involvement, and so on). Equations (1) and (2) do not include the year and grade FE typically included in VAMs because (1) can only be estimated for one cross section of kindergarteners and (2) can only be estimated for one cross section of first graders. We stress that these specifications identify classroom, as opposed to teacher, effects because the cohort nature of the ECLS-K data contains only one observation per teacher and classroom effects are treated as fixed rather than random. For example, we cannot distinguish teacher effects from class size effects and thus focus on estimating classroom effects on students' achievement. To examine the ability of the basic demographic variables typically observed in administrative data, and that of the rich summer activity variables observed in the ECLS-K, to control for summer learning, we estimate equations (1) and (2) using various specifications of **x**.

We take Ordinary Least Squares (OLS) estimates of (1) and (2) as the baseline for three reasons. First, most existing consequential accountability policies employ similar lag-score specifications (Papay, 2011). Second, the prevailing consensus among researchers is that this relatively straightforward approach likely outperforms more sophisticated models and estimation strategies, as most sorting of students to classrooms is based on lagged achievement. For example, simulation evidence from Guarino, Reckase, and Wooldridge (2015) find "dynamic OLS" estimates to be the most robust to a variety of potential non-random student-teacher

assignment scenarios. Third, the ECLS-K only spans one summer vacation, so alternative estimators that require additional years of data are infeasible (e.g., first-differenced instrumental variable procedures that require multiple lags of student achievement to use as instruments). Nonetheless, we conduct a series of sensitivity analyses including estimating gain-score specifications, relaxing the "five student per classroom" sample restriction, and estimating un-weighted regressions to examine the robustness of the main results.

After estimating equations (1) and (2) we examine the stability of the estimated classroom effects by comparing rankings of the estimated classroom effects in (1a) to those in (1b), and similarly for (2a) and (2b). Intuitively, only the classroom effect estimates in (1b) and (2b) are potentially biased by the non-random distribution of summer learning loss across classrooms. We focus on rankings of classroom effects rather than the estimated FE themselves, as VAMs frequently produce reliable rankings of teacher effectiveness even when the point estimates are inconsistent or imprecisely estimated (Guarino, Reckase, Stacy, & Wooldridge, 2015) and valid rankings are arguably more policy relevant than point estimates of teachers' effectiveness.

We compare the rankings generated by cross- and within-year VAMs in two ways that are similar to the ways in which previous researchers have compared rankings generated by different VAM specifications (e.g., Guarino Reckase, & Wooldridge, 2015; Harris et al., 2012; Loeb & Candelaria, 2012; Koedel & Betts, 2007; McCaffrey et al., 2009). First, we estimate the Spearman Rank Correlation Coefficient, which is a simple summary statistic of the similarity between two rankings. Second, we construct transition matrixes that report the frequency and type of classrooms' quintile-rank switching across specifications, which provide a more nuanced understanding of how the rankings change and of the implications for policies that penalize

(reward) teachers in the bottom (top) of the effectiveness distribution.

Results

Table 3 reports Spearman rank correlations of the comparisons between estimated kindergarten classroom effects generated by equations (1a) and (1b) and between estimated first grade classroom effects generated by equations (2a) and (2b) for the baseline specification as well as several alternative specifications. The Spearman rank correlations suggest that estimated classroom effects on math achievement are less robust to test timing than those on reading achievement, though for kindergarten classrooms neither is particularly stable. The greater robustness of the reading rankings might be due to some combination of the relatively higher reliability of the ECLS-K reading assessments and the slightly greater variation in math summer learning rates, though it is impossible to identify the exact causes given the available data; it would be interesting to see if this pattern holds in similar analyses of other district or state administrative data.[4]

Interestingly, these results are quite robust to the choice of statistical controls in **x**, functional form of the test scores, weighting, and sampling restrictions for both subjects. Specifically, controlling for observed student characteristics and summer activities does not change the estimated Spearman rank correlation for either subject. The inability of conditioning on children's summer activities to improve the validity of the cross-year VAM estimates likely results from some combination of the general findings that conditioning on lagged achievement alone is enough to obtain unbiased estimates of teacher effectiveness (Chetty et al., 2014; Kane & Staiger, 2008), that only about 10% of the variation in summer learning rates can be explained by observed student and household characteristics (Downey et al., 2004), and that the ECLS-K

information on summer activities are not particularly detailed with regards to the quantity or quality of summer programs and parental involvement. Gain-score VAMs that restrict α to equal one perform slightly better than the baseline lag-score specification for math achievement, as the Spearman rank correlation is about 12 percentage points larger than that for the baseline model, and this increase in stability is entirely concentrated in the bottom half of the effectiveness distribution. However, there are no such differences between the lag- and gain-score models for reading achievement, which is consistent with Quinn's (2015) analysis of racial differences in average summer learning rates.

The findings for both subjects remain qualitatively similar across the remaining sensitivity analyses. Specifically, these include estimating the baseline models using either unstandardized (raw) vertically scaled test scores or theta-score estimates of students' latent ability, estimating un-weighted baseline models that do not adjust for the ECLS-K's nonrandom sampling frame, relaxing the classroom-size sample restriction from five students per classroom to three students per classroom but only ranking classrooms with at least five students, and relaxing the classroom-size sample restriction from five students per classroom to three students per classroom and ranking all classrooms. The similarity between the weighted (baseline) and un-weighted results suggest that the VAMs are correctly specified and that the main results are not driven by the potential drop in precision associated with the use of sampling weights (Solon et al., 2015). Similarly, the main results' robustness to the students-per-classroom restriction suggests that the findings are not driven by imprecision associated with the relatively small number of students per classroom.

The correlations reported for first grade classrooms in table 3 are qualitatively similar to those for kindergarten classrooms in that the year-to-year and fall-to-spring estimates are

positively but not perfectly correlated and classroom-effect rankings for reading are more robust to test timing than those for math. However, a notable difference between the fall-fall and spring-spring analyses is that the spring-spring results for first-grade classrooms are more stable for both math and reading: the Spearman correlations are between 0.8 and 0.9. As in the analysis of kindergarten classrooms reported in table 3, the first-grade results are remarkably robust to a variety of alternative specifications, assessments, weighting schemes, and sample restrictions.

The only sensitivity analysis to yield a substantive difference from the baseline results is the gain score model, which yields less stable results for both subjects. One possible explanation of this difference is that the mechanism by which students are assigned to classrooms in kindergarten is different from the mechanism used in first grade. Specifically, classroom assignments in first grade may rely more on students' academic ability and past performance, as such information is more readily available to school administrators for students entering first grade than for students entering kindergarten. If this is true, conditioning on lagged achievement plays a larger role in identifying valid estimates of first-grade classroom effects than kindergarten classroom effects (e.g., Chetty et al., 2014; Quinn, 2015).

While the correlations presented in table 3 indicate substantive differences between the classroom-effectiveness rankings generated by fall-to-spring and fall-to-fall achievement gains, particularly in math, transition matrixes that report the movement of classrooms across quintiles of the classroom-effectiveness distribution provide a more nuanced understanding of the stability of such rankings. Table 4 presents two such transition matrixes for math and reading achievement based on the baseline VAM that conditions on the elements of **x** typically observed in administrative data. Like in table 3, the transition matrixes for the alternative specifications considered above are qualitatively similar and thus not reported in the interest of brevity. The

diagonal elements of the transition matrixes reported in table 4 represent classrooms that were in the same quintile of the effectiveness rankings generated by fall-to-fall and fall-to-spring VAMs. As expected given the results in table 3, the figures along the diagonals are significantly lower than 100%, reinforcing the general finding that kindergarten classroom effectiveness rankings are sensitive to the timing of the assessments used in the VAM. Indeed, only about half of classrooms ranked in the lowest or highest quintiles of math effectiveness remained in the same quintile in both the within-year and cross-year rankings. Furthermore, nearly ten percent of these classrooms experienced large swings from the highest to lowest quintile, or vice versa, depending on whether a within-year or cross-year VAM was estimated. As suggested by the higher correlations for reading in table 3, the transition matrix for reading VAMs reported in table 4 suggests significantly more stability: About 60% of bottom-quintile classrooms and 70% of top-quintile classrooms remain in the same quintile regardless of whether cross-year or within-year test scores are used and large changes between the top and bottom quintiles are nonexistent. Together, the results presented in tables 3 and 4 suggest that a nontrivial subset of teachers may be misclassified by evaluations that rely on cross-year fall-fall VAMs, and that such misclassifications are both more frequent and larger in magnitude with regards to math effectiveness than reading effectiveness.

Table 5 replicates the transition matrix analysis in table 4 for specifications (2a) and (2b), comparing rankings of first-grade classrooms generated by within-year fall-to-spring VAMs to those generated by cross-year spring-to-spring VAMs. The transition matrixes reported in table 5 show that the spring-spring VAMs are more stable than the fall-fall VAMS, particularly for reading, and large swings across multiple quintiles are exceedingly rare in both subjects. Still, 20% to 35% of teachers who are in either the top or bottom quintile in one ranking do not remain

in the same quintile when the timing of the lagged test score is changed. The results presented in table 5 suggest that a nontrivial subset of teachers may be misclassified by evaluations that rely on cross-year spring-spring VAMs, that such misclassifications are both more frequent and larger in magnitude with regards to math effectiveness than reading effectiveness, and that such misclassifications are both less frequent and smaller in magnitude than analogous misclassifications associated with the cross-year fall-fall VAMs discussed in table 4.

Discussion and Policy Implications

The majority of current test based-accountability and teacher-evaluation programs that rely on teacher, classroom, or school value-added measures compute value-added by measuring achievement gains from one academic year to the next. For example, the Houston and Nashville school districts administer standardized tests each spring that measure students' achievement gains between the spring of grade *g*-1 and the spring of grade *g*. This is potentially problematic, as students' summer gains and losses are incorrectly attributed to students' grade-*g* teachers and schools. Indeed, previous research has discussed the potential bias in VAM-based estimates of school effectiveness caused by the use of "cross-year" or spring-to-spring achievement gains (Downey et al., 2008; McEachin & Atteberry, 2014). However, the practical importance of this bias in the context of VAM-based measures of teacher effectiveness is unknown.

The current study contributes to this gap in the literature by providing evidence on the validity of value-added estimates of classroom effects generated by fall-to-fall and spring-to-spring "cross-year" VAMs relative to arguably more valid fall-to-spring "within-year" VAMs. We consistently find that estimated classroom effects are unstable between "cross-year" and "within-year" VAMs. Specifically, only 50% of kindergarten classrooms ranked in the lowest

quintile by a fall-to-spring math achievement VAM remain in the lowest quintile of a fall-to-fall VAM, and about 10% of those classrooms move to the highest quintile. These results suggest that policies that reward teachers based on their position in the distribution of teacher effects estimated by "cross-year" VAMs, such as Nashville's Project on Incentives in Teaching (POINT) program (Spring et al., 2011), misclassify a nontrivial fraction of teachers.

Similarly, about 65% of first-grade classrooms ranked in the lowest quintile by a fall-to-spring math achievement VAM remain in the lowest quintile of a spring-to-spring VAM, though extreme swings between the top and bottom quintiles are exceedingly rare among first-grade teachers. This may suggest that spring testing is preferable to fall testing, though we cannot rule out the possibility that estimates of first-grade classroom effects are inherently more stable than those of kindergarten classrooms. Our results are largely consistent with those of school-level VAM analyses, which also stress the importance of test timing in the context of VAM-based accountability schemes. However, because the current study is limited to one cohort of students and only observes teachers during one academic year, future work that conducts similar analyses of longitudinal administrative data spanning multiple cohorts of students who were administered similar fall and spring tests will prove fruitful. Finally, an interesting non-finding of the current study is that conditioning on students' characteristics and summer activities in "cross-year" VAMs does not significantly improve the stability of estimated classroom effects.

To place these results in the context of the broader literature on the stability of value-added estimates of teacher effectiveness, it is instructive to compare the estimated rank correlations between "cross-year" and "within-year" VAMs to those between different years and subjects in the existing literature. The fall-fall math and reading correlations of 0.5 and 0.8, respectively, fall at the high end of estimated intertemporal correlations. The spring-spring math

and reading correlations of 0.8 and 0.9 are larger. For example, McCaffrey et al. (2009) find intertemporal correlations as large as 0.6 and Goldhaber and Hansen (2013) find correlations between 0.6 and 0.8 for math and between 0.5 and 0.7 for reading. The "cross-year" and "within-year" rankings estimated in the current study are also more stable than those across subjects, as studies of the stability of teacher rankings across math and reading find rank correlations of about 0.6 (Koedel & Betts, 2007; Loeb et al., 2012). However, despite the positive and sometimes relatively high rank correlation coefficients estimated in the current study, some critics have argued that rank correlations of these magnitudes indicate that VAM-based estimates of teacher effectiveness are invalid and thus should not be used to make high-stakes personnel decisions (Hill, 2009).

While the ECLS-K data provide a unique opportunity to evaluate the practical implications of summer learning loss for VAM-based estimates of classroom effectiveness using nationally representative data and provide rich information on students' summer activities, these data are not without limitations. Specifically, three limitations are worth mentioning in the hope that future research might address these shortcomings using different data. First, as mentioned in the Methodology section, the fact that the ECLS-K followed one cohort of children means that teachers are only observed at one time point. This limits the types of analyses that can be conducted; for example, ECLS-K data cannot be used to examine the implications of summer learning loss for estimating the intertemporal stability of VAM-based estimates of teacher effectiveness. Second, because the ECLS-K sampled students within classrooms, only a small number of student observations are available for certain classrooms, which limits the precision of estimated classroom effects. Finally, while the ECLS-K is a nationally representative sample of the 1998-99 U.S. kindergarten cohort, it is not necessarily a representative sample of the teachers

who are subject to high-stakes accountability programs or evaluated by VAM-based measures of effectiveness. This is because most such policies and programs in the U.S. target grades three through eight. For these reasons, conducting similar analyses using state- or district-level administrative data would usefully further our understanding of how summer learning affects the validity of VAM-based measures of teacher effectiveness.

Taken at face value, the results of the current study suggest that there are benefits to testing students twice per year. In addition to providing arguably more accurate estimates of teacher and school effectiveness, a biannual fall-spring testing regime would provide teachers and school administrators with accurate and current information on students' achievement at the start of the school year. For example, such information could be used to design review sessions and lesson plans early in the fall semester. Similarly, these data could be used to identify the students who experienced the greatest levels of summer learning loss and target high-quality summer programs to such students in subsequent summers. Building Educated Leaders for Life (BELL) and KindergARTen are two examples of full-day six-week summer programs that are known to improve academic achievement (Borman et al., 2009; Chaplin & Capizzano, 2006).

Biannual fall-spring testing is not costless, of course, for a variety of reasons. First, if some teachers have persistent effects on student achievement that operate through effects on students' summer activities or parental involvement, fall-spring VAMs would fail to capture this dimension of teacher effectiveness (von Hippel, 2009). Second, there are explicit costs in terms of both time and money associated with doubling the number of tests. Downey et al. (2008) rightly note that in the case of school evaluation, within-year VAMs could be implemented without doubling the number of tests simply by shifting every other spring test to the following fall. However, in the context of teacher evaluation such a reshuffling would further reduce the

number of teachers teaching in "tested grades" who can be evaluated by VAM-based measures. Finally, it is possible that using within-year fall-spring VAM scores to evaluate teachers would encourage teachers to game the system by artificially depressing fall scores (Baker et al., 2010) and effectively penalize teachers who create achievement gains prior to the fall baseline score when the fall baseline is not administered early enough in the school year. If teachers reacted in this way, measures of both VAM-based teacher effectiveness and students' summer learning would be systematically biased.

As a result of these concerns and related political considerations, biannual fall-spring testing is unlikely to be instituted on a large scale in the immediate future. Nonetheless, the instability between "cross-year" and "within-year" teacher VAMs documented in the current study has at least four implications for current education policy and practice. First, these findings suggest that in addition to the statistical properties of VAM specifications and estimators; the sorting of students and teachers into schools, tracks, and classrooms; and the content of student assessments; researchers, educators, and policymakers should pay greater attention to the timing of the assessments used by VAM and related effectiveness measures. More generally, these results highlight the importance of summer learning in educational policy and practice, regarding both how summer learning complicates evaluation of the efficacy of educational interventions and how schools and teachers might influence students' summer learning.

Second, the robust finding that controlling for the summer activities observed in the ECLS-K does not noticeably improve the stability of effectiveness rankings suggests that collecting such information is unlikely to be a good use of school districts' scarce resources. This result is not entirely surprising, as Downey et al. (2004) find in the ECLS-K that only about 10% of the variation in summer learning rates can be explained by observed student and household

characteristics. However, the inability of the ECLS-K's information on students' summer activities to either predict summer learning or improve the performance of cross-year VAMs may result from the imprecise nature of the summer activity variables; for example, the ECLS-K summer activity variables do not account for the quality or academic orientation of summer programs. Accordingly, it would be useful for future research to verify this non-finding in other contexts, using data on participation in other types of summer activities, and using data on summer programs' focus and quality.

Third, the greater stability of the spring-spring first-grade rankings suggests that when "cross-year" VAMs are necessary due to data or resource limitations, conducting annual assessments in the spring may be preferable to doing so in the fall. An important caveat to this result is that the spring-spring and fall-fall analyses of the ECLS-K data were necessarily conducted on first-grade and kindergarten classrooms, respectively, and the result could be driven by grade-specific differences in either the composition of the teaching force or in the way students are sorted into classrooms. Again, it would be useful to further investigate the relative stability of fall-fall versus spring-spring "cross-year" VAMs using data that permits both analyses of the same teachers and students. Specifically, to analyze teachers in grade g, this would require data on four assessments: spring of g-1, fall and spring of g, and fall of g+1.

Finally, the current study reinforces the importance of using multiple measures to evaluate teachers (e.g., Polikoff & Porter, 2014). There are tradeoffs associated with each approach to measuring teacher effectiveness and no one measure is perfect. Specifically, when cross-year (e.g., spring-to-spring) VAM scores are used as a measure of teacher effectiveness, it is perhaps particularly important to augment such measures with alternatives that are not sensitive to variation in students' summer learning rates.

Notes

1. This, and all subsequent sample sizes, are rounded to nearest 50 in accordance with NCES regulations for the use of restricted ECLS-K data.

2. Theta test scores and theta coefficients of internal consistency are different constructs. The former capture student-specific aptitude in the academic skills measured by the ECLS-K assessments. The latter measure each test's reliability, or extent to which rank orders of student performance would be preserved if students took the same test multiple times (with zero memory of previous attempts).

3. Specifically, we proceed in two steps. First, we follow Burkam, Ready, Lee, and LoGerfo (2004) in regressing the "summer gain score" on a constant, the number of school days that occurred after the spring-kindergarten test, and the number of school days that occurred before the fall-first grade test. We then compute each student's "true" summer gain (loss) by subtracting the estimated contribution of school days from the total gain (loss) experienced between the spring-kindergarten and fall-first grade tests. The average of these "true" summer learning estimates equals the estimated intercept from the regression in step 1.

4. The theta coefficients of internal consistency for the four math are between 0.92 and 0.94 while the corresponding reading coefficients are between 0.93 and 0.97 (Rock & Pollack, 2002).

References

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, 23, 171-191.

Armour-Garb, A. (2009). Should "value-added" models be used to evaluate teachers? *Journal of Policy Analysis and Management*, 28, 692-693.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, 4, 351-383.

Becker, G., & Tomes, N. (1986). Human capital and the rise and fall of families. *Journal of Labor Economics, 4*, S1-S39.

Borman, G. D., Benson, J., & Overman, L. T. (2005). Families, schools, and summer learning. *The Elementary School Journal*, 106, 131-150.

Borman, G. D., & Boulay, M. E. (2004). *Summer learning: Research, policies, and programs*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Borman, G. D., Goetz, M. E., & Dowling, N. M. (2009). Halting the summer achievement slide: A randomized field trial of the KindergARTen Summer Camp. *Journal of Education for Students Placed at Risk*, 14(2), 133–147.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Measuring Test Measurement Error A General Approach. *Journal of Educational and Behavioral Statistics, 38*(6), 629-663.

Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education*, 77, 1-31.

Chaplin, D., & Capizzano, J. (2006). Impacts of a summer learning program: A random assignment study of Building Educated Leaders for Life (BELL). Washington, DC: Urban Institute.

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

Chin, T., & Phillips, M. (2004). Social reproduction and child-rearing practices: Social class, children's agency, and the summer activity gap. *Sociology of Education*, 77, 185-210.

Cooper, H., Charlton, K., Valentine, J. C., & Muhlenbruck, L. (2000). Making the most of summer school: A meta-analytic and narrative review. *Monographs of the Society for Research in Child Development*, 65, 1-127.

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, 227-268.

Covay, E., & Carbonaro, W. (2010). After the bell: Participation in extracurricular activities, classroom behavior, and academic achievement. *Sociology of Education*, 83, 20-45.

Currie, J., & Thomas, D. (2001). Early test scores, socioeconomic status, school quality, and future outcomes. *Research in Labor Economics*, 20, 103-132.

Dieterle, S. G., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). How do principals assign students to teachers? Finding evidence in administrative data and the implications for value-added. *Journal of Policy Analysis & Management*, 34(1), 32-58.

Downey, D. B., von Hippel, P. T., & Broh, B. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, 69, 613-635.

Downey, D. B., von Hippel, P. T., & Hughes, M. (2008). Are "failing" schools really failing? Using seasonal comparison to evaluate school effectiveness. *Sociology of Education*, 81(3), 242-270.

Entwisle, D. R., & Alexander, K. L. (1992). Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review*, 57, 72-84.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, School, and Inequality*. Boulder, CO: Westview.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2001). Keep the faucet flowing: Summer learning and home environment. *American Educator*, 25, 10-15, 47.

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, 269-279.

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

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating Teachers: The Important Role of Value-Added*. Brookings Institution. Unpublished manuscript.

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

Guarino, C. M., Reckase, F., Stacy, B. W., & Wooldridge, J. M. (2015). Evaluating specification tests in the context of value-added estimation. *Journal of Research on Educational Effectiveness*, 8(1), 35-59.

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

Guryan, J., Hurst, E., & Kearney, M. (2008). Parental education and parental time with children. *The Journal of Economic Perspectives*, 22, 23-46.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100(2), 267-271.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management*, 28, 693-699.

Harris, D. N. (2011). *Value-added measures in education*. Cambridge, MA: Harvard Education Press.

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

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, 28, 700-709.

Hoover-Dempsey, K. V., & Sandler, H. M. (1995). Parental involvement in children's education: Why does it make a difference? *Teachers College Record*, 97, 310-331.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89, 761-796.

Kane, T. J., and D.O. Staiger. 2008. *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.

Kaushal, N., Magnuson, K., & Waldfogel, J. (2011). How is family income related to investments in children's learning? In G. Duncan & R. Murnane (Eds.). Whither opportunity? Rising inequality, schools, and children's life chances (pp. 187-206). New York, NY: Russell Sage Foundation.

Kelly, S. (2012). Understanding teacher effects: Market versus process models of educational improvement. In S. Kelly (Ed.), *Assessing Teacher Quality: Understanding Teacher Effects on Instruction and Achievement* (pp. 7-32). New York, NY: Teachers College Press.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function* Vanderbilt, Peabody College. Unpublished manuscript.

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

Lareau, A. (2003). *Unequal childhoods: Class, race, and family life*. Berkeley, CA: University of California Press.

Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. W. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications* (pp. 195-212). Maple Grove, MN: JAM Press.

Loeb, S., & Candelaria, C. (2012). *How stable are value-added estimates across years, subjects, and student groups?* The Carnegie Knowledge Network. Unpublished manuscript.

Loeb, S., Kalogrides, D., & Beteille, T. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy,* 7, 269-304.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: RAND.

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

McEachin, A. & Atteberry, A. (2014) The Impact of Summer Learning Loss on Measures of School Performance. *EdPolicyWorks Working Paper Series, No. 26.*

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26, 237-257.

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

Phillips, M. (2011). Parenting, time use, and disparities in academic outcomes. In G. Duncan & R. Murnane (Eds.). Whither opportunity? Rising inequality, schools, and children's life chances (pp. 207–228). New York, NY: Russell Sage Foundation.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, 36(4), 399-416.

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

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73, 417-458.

Rock, D. A., & Pollack, J. M. (2002). Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K): Psychometric Report for Kindergarten through First Grade. NCES Working Paper 2002-05.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125, 175-214.

Sass, T. R., Harris, D. N., & Semykina, A. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, 38, 9-23.

Solon, G., S. J. Haider, & J. M. Wooldridge. (2015). What are we weighting for? *Journal of Human Resources*, 50(2), 301-316.

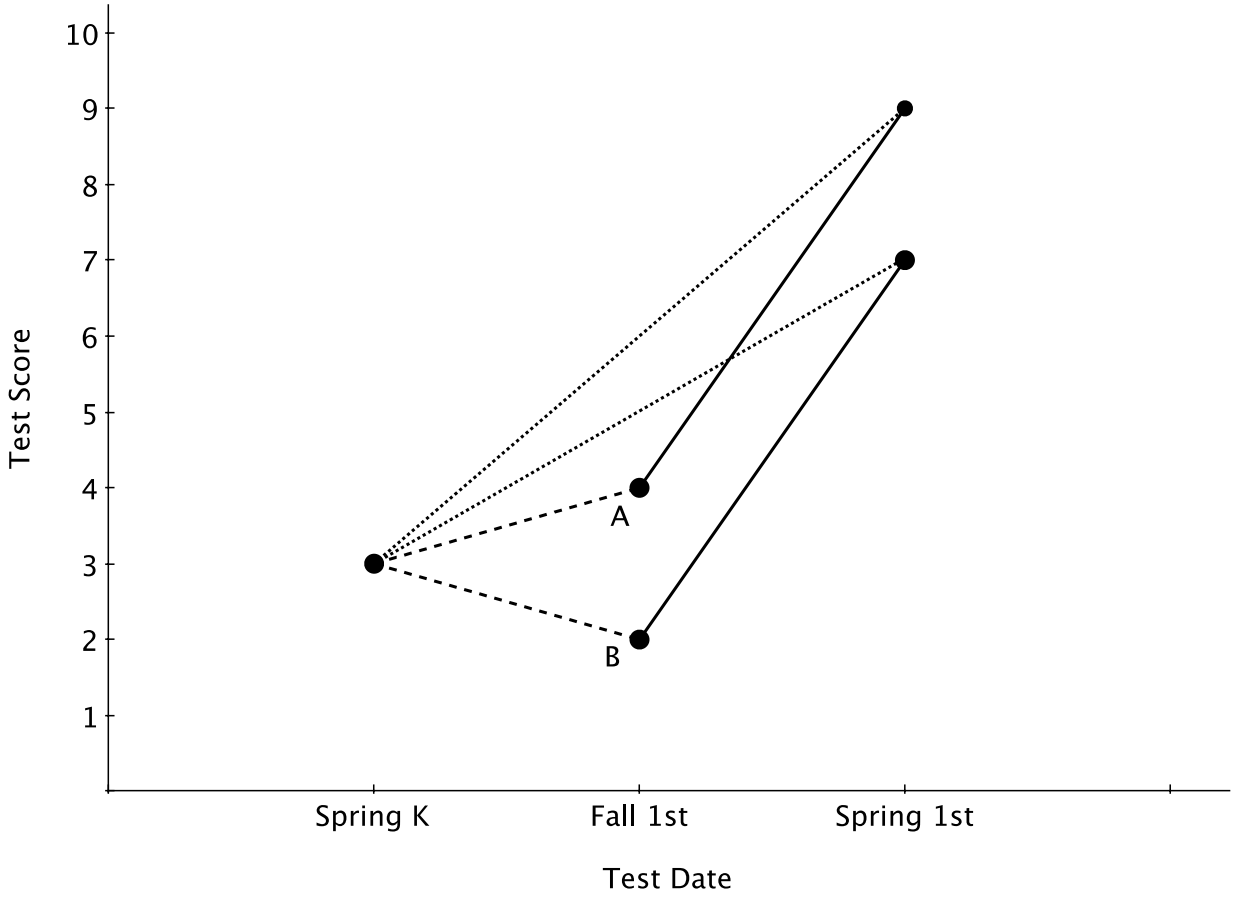
Springer, M. G., Ballou, D., Hamilton, L. S., Le, V. N., Lockwood, J. R., McCaffrey, D. F., et al. (2011). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT)*. Society for Research on Educational Effectiveness.

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

von Hippel, P. T. (2009). Achievement, learning, and seasonal impact as measures of school effectiveness: It's better to be valid than reliable. *School Effectiveness and School Improvement*, 20(2), 187-213.

Winters, M. A., & Cowen, J. M. (2013). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. *Journal of Policy Analysis and Management*, 32, 634-654.

Figure 1. How Summer Learning Impacts Cross-Year Value-Added (VA) Estimates



Notes: Dashed lines represent summer learning. Solid lines represent school-year learning. Dotted lines represent cross-year "spring-to-spring" learning.

Figure 2. Distribution of Summer Achievement Gains

Figure 2A. Distribution of Summer Math Gains

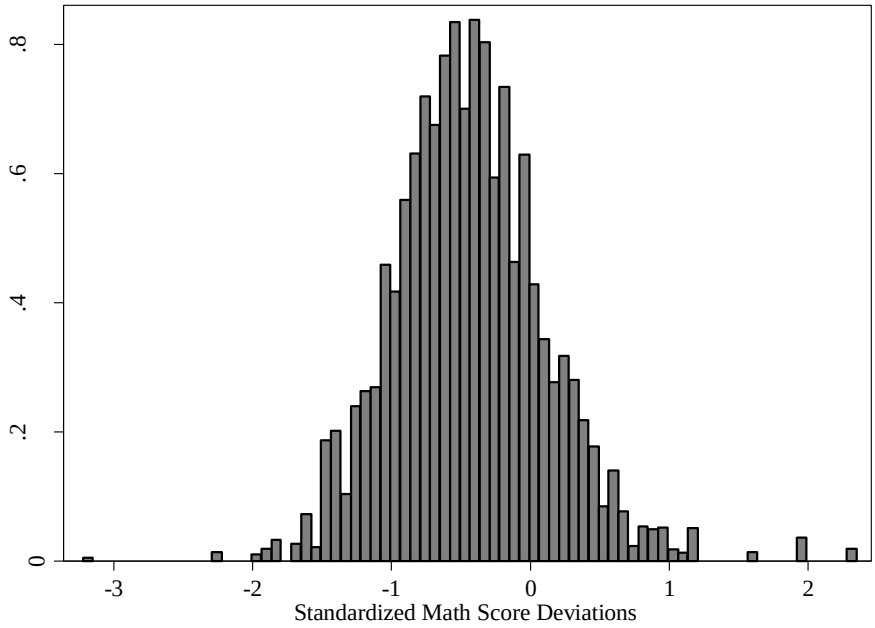
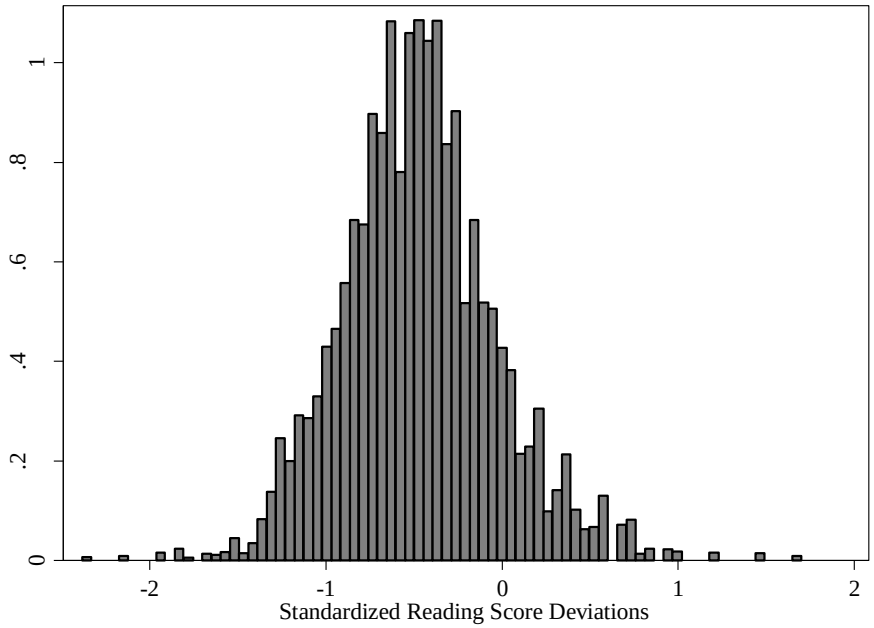


Figure 2B. Distribution of Summer Reading Gains



Notes: The histograms in figure 2 are weighted to adjust for unequal probabilities of sample selection. The gain scores are adjusted to account for the timing of assessments.

Figure 3. Distribution of Classroom-Level Summer Learning Gains

Figure 3.A Mean Math Gains

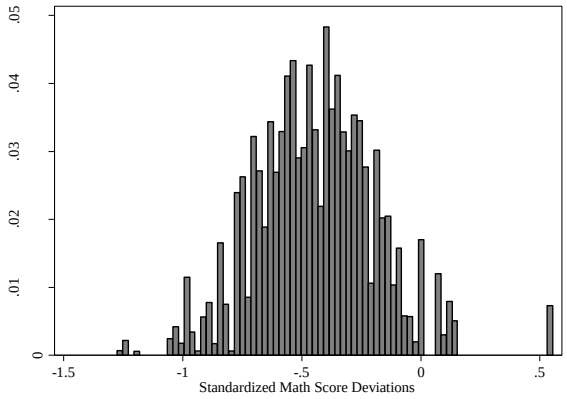


Figure 3.B Median Math Gains

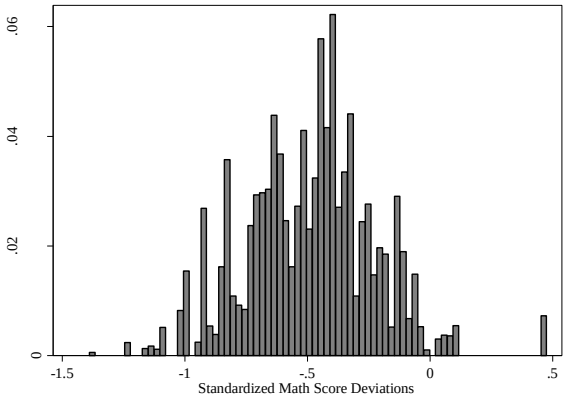


Figure 3.C Mean Reading Gains

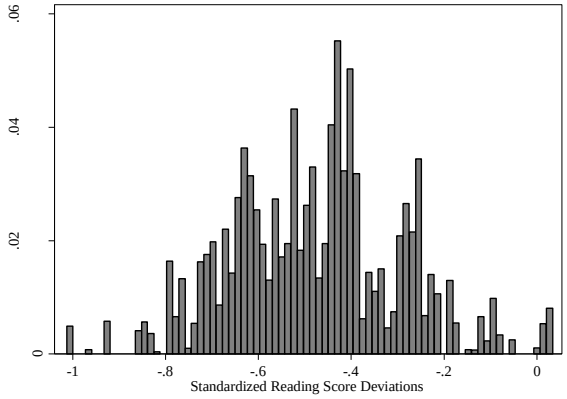
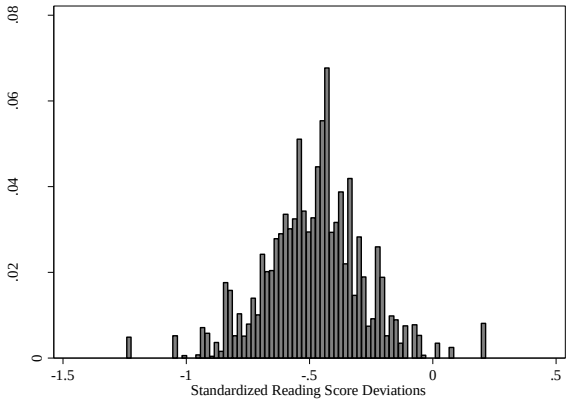


Figure 3.D Median Reading Gains



Notes: The histograms in figure 3 are weighted to adjust for unequal probabilities of sample selection. The gain scores are adjusted to account for the timing of assessments.

Table 1: Descriptive Statistics of Summer Learning

	Mean	S.D.
Math Summer Learning Gain	-0.43	0.56
Within-School S.D.		0.50
Within-Classroom S.D.		0.49
Reading Summer Learning Gain	-0.48	0.44
Within-School S.D.		0.41
Within-Classroom S.D.		0.40
Kindergarten Students	1,500	
First-grade Students	1,250	
Kindergarten Classrooms	200	
First-grade Classrooms	150	
Schools	100	

Notes: All estimates are weighted to account for the unequal probabilities of sample selection by NCES-provided sampling weights. Means and SD are computed for the 1,500 students for whom summer learning gains (losses) can be computed. Sample sizes are rounded to the nearest 50, in accordance with NCES regulations for restricted-use ECLS-K data. Summer learning gains are adjusted for the timing of both spring and fall tests and are normalized to have mean zero and standard deviation one.

Table 2: Descriptive Statistics of Student Characteristics and Summer Activities

	Mean	S.D.
Student Characteristics		
White	79.4%	
Black	6.6%	
Hispanic	7.7%	
Other race/ethnicity	6.2%	
Female	51.1%	
Poverty	10.3%	
Does not speak English at Home	2.9%	
Has Individualized Education Plan (IEP)	4.3%	
Kindergarten Redshirt	7.9%	
Attends Private School	29.5%	
Attends Urban School	28.9%	
Attends Suburban School	36.2%	
Attends Rural School	34.8%	
Mom No H.S. Degree	5.5%	
Mom H.S. Degree	33.2%	
Mom Some College	30.4%	
Mom Bachelor's Degree or more	30.9%	
Summer Activities		
Organized summer activities	60.6%	
Attended summer school	10.2%	
# of trips to library/bookstore	7.2	7.1
Child never practice math	19.3%	
Child sometimes practices math	72.0%	
Child practices math everyday	8.7%	
Mother never reads to child	2.5%	
Mother sometimes reads to child	50.2%	
Mother reads to child everyday	47.3%	
N Children	1,500	
N Schools	100	

Notes: Means and standard deviations (SD) are weighted by NCES provided sampling weights to account for unequal probabilities of sample selection. SD are only reported for non-binary variables.

Table 3: Spearman Correlation Coefficients for Ranking Comparisons

	Kindergarten Classrooms Fall-Spring vs. Fall-Fall Gains	First-Grade Classrooms Fall-Spring vs. Spring-Spring
Math Achievement		
No controls	0.45	0.79
Baseline	0.45	0.77
Rich control set	0.45	0.80
Gain Score VAM	0.57	0.66
Baseline, un-standardized	0.43	0.79
Baseline, Theta test scores	0.39	0.78
Baseline, un-weighted	0.46	0.81
Baseline, relaxed sample	0.49	0.80
Baseline, relaxed sample and all classrooms	0.46	0.80
Reading Achievement		
No controls	0.80	0.92
Baseline	0.80	0.93
Rich control set	0.80	0.92
Gain Score VAM	0.81	0.84
Baseline, un-standardized	0.78	0.92
Baseline, Theta test scores	0.79	0.87
Baseline, un-weighted	0.78	0.93
Baseline, relaxed sample 1	0.81	0.91
Baseline, relaxed sample 2	0.78	0.90
Students	1,500	1,250
Classrooms	200	150
Students (sample 1)	2,100	1,900
Classrooms (sample 1)	200	150
Students (sample 2)	2,100	1,900
Classrooms (sample 2)	400	400

Notes: The correlation coefficients reported for kindergarten (first grade) in this table compare rankings of classroom effects generated by equations 1a and 1b (2a and 2b) of the main text.

Table 4: Quintile Transitions of Kindergarten Classroom Effects Generated by Fall-Spring and Fall-Fall Gains

	Fall-fall, baseline model				
	Math Achievement				
Fall-Spring	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Quintile 1	55.6	11.1	13.9	11.1	8.3
Quintile 2	11.1	30.6	22.2	27.8	8.3
Quintile 3	8.3	30.6	27.8	19.4	13.9
Quintile 4	16.7	16.7	22.2	19.4	25.0
Quintile 5	8.3	11.1	13.9	22.2	44.4
	Reading Achievement				
Fall-Spring	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Quintile 1	61.1	19.4	16.7	2.8	0.0
Quintile 2	25.0	41.7	19.4	11.1	2.8
Quintile 3	11.1	30.6	36.1	22.2	0.0
Quintile 4	2.8	5.6	22.2	44.4	25.0
Quintile 5	0.0	2.8	5.6	19.4	72.2

Notes: The statistics reported in this table compare rankings of classroom effects generated by equations 1a and 1b of the main text. The sample contains 1,500 students in 200 classrooms.

Table 5: Quintile Transitions of First-Grade Classroom Effects Generated by Fall-Spring and Spring-Spring Gains

	Fall-fall, baseline model				
	Math Achievement				
Fall-Spring	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Quintile 1	64.7	23.5	8.8	0.0	2.9
Quintile 2	24.2	39.4	27.3	9.1	0.0
Quintile 3	11.8	29.4	23.5	23.5	11.8
Quintile 4	0.0	6.1	24.2	48.5	21.2
Quintile 5	0.0	0.01	18.2	18.2	63.6
	Reading Achievement				
Fall-Spring	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Quintile 1	79.4	17.6	2.9	0.0	0.0
Quintile 2	21.2	54.5	24.2	0.0	0.0
Quintile 3	0.0	20.6	44.1	32.4	2.9
Quintile 4	0.0	6.1	30.3	42.4	21.2
Quintile 5	0.0	0.0	0.0	24.2	75.8

Notes: The statistics reported in this table compare rankings of classroom effects generated by equations 2a and 2b of the main text. The sample contains 1,250 students in 150 classrooms.