The Role of Summer Misattribution in Estimation of Teacher Value-Added

Allison Atteberry, PhD
Assistant Professor,
Research & Evaluation Methods
CU-Boulder School of Education
allison.atteberry@colorado.edu

*Daniel Mangan**
**Doctoral Student*
Educational Foundations, Policy, & Practice
CU-Boulder School of Education
daniel.mangan@colorado.edu

We would like to thank John Papay for his careful read of drafts of this manuscript, given its close correspondence with his own work. The views expressed in the article are solely those of the authors. Any errors are attributable to the authors.

Introduction

Over the last ten years, the majority of U.S. states have revised their teacher evaluation policies, largely stimulated by the federal *Race to the Top* (RTTT) grant initiative. A key component of *RTTT* encouraged states to more tightly link teacher evaluations to student outcomes. This stipulation was often met by incorporating value-added measures (aka VAM scores). This class of statistical models attempts to isolate the causal effect of individual teachers on student achievement typically by leveraging a fixed effects framework, covariate adjustment, and—importantly—controlling for a student's prior performance in the previous spring. Yet estimating causal effects of teachers in this manner is far from straightforward; it is unclear that any measures produced in this way are sufficiently unbiased and precise to use in high-stakes decisions about teacher hiring, retention, compensation, professional development, and tenure. If VAM scores do not reflect teachers' true impacts on students, then putting them to policy-use threatens to both mislead policymakers and undermine teachers' trust commitment to the profession.

In a 2011 study of the validity of VAM scores to capture teacher effectiveness, Papay briefly shows in his final table that scores generated by models using the traditional last-spring to current-spring test timings are essentially orthogonal (correlation of -0.10) to those same teachers' estimates when constructed using current-fall to next-fall[1] test timings. Though not the focus of his paper, this finding should be of great concern: There is no principled reason to use spring-to-spring ("SS")[2] over fall-to-fall ("FS") test timings to construct these measures, and according to

[1] Papay is able to examine this because, as in the current study, he had access to more than one test, including one that was administered in both the fall and spring of every school year.

[2] "SS" refers to the typical "prior-spring" to "current-spring" annual testing administration schedule used by the vast majority of U.S. states. "FS" refers to a more desirable but unlikely scenario in which students'

Papay's results, this choice—made solely as an artifact of the timing of statewide testing systems—would lead to an entirely different ranking of teachers' effectiveness. However, since test timing was not the focus of the Papay paper, it is unclear whether the finding would hold up on other tests or according to other specifications of the models used to produce the VAM scores. Yet this little-noticed finding embedded in that 2011 Papay study could have serious policy implications and warrants further investigation.

We now replicate this finding in another school district using panel data from 2011-12 to 2014-15. Because the district from which our data is drawn administers one of its district-wide assessments in both fall and spring, we are able to do what is typically not possible with statewide standardized test score data: We estimate each teacher's VAM score using spring-to-spring ("SS"), fall-to-fall ("FF"), and fall-to-spring ("FS") test administration timings. Our finding is consistent with Papay's: There is little to no correlation between SS-based and FF-based VAM scores (0.07 for ELA, -0.02 for math). This finding—now apparent in two different educational contexts—raises concerns about the increasing prevalence of school districts' use of teacher value-added estimates in high-stakes policy decisions.

Framing the Problem

Why could test timing potentially have a significant effect on teacher VAM scores? For a statistical model designed to isolate causal effects of teachers on student test scores during the school year, it would be ideal to administer a highly reliable assessment at the very start and very

achievement gains in a given year are measured specifically from the current fall to the current spring of that school year. "FF" refers to a scenario in which achievement gains are estimated from a "current-fall" to "following-fall" ("FF") testing administration schedule.

end of each school year; this “FS” timing would most clearly identify a student’s achievement gains that occurred while assigned to a given teacher. A necessary artifact of only administering standardized tests to students once per year is that a summer period will be inappropriately attributed to the teacher, even though the teacher does not interact with the student during that time. Under the typical “SS” test administration schedule, the summer before the teacher encounters the student is incorporated into their VAM score. Under a hypothetical “FF” administration schedule, the summer *after* the teacher is assigned to the student is incorporated to their annual VAM score. See Figure 1, modified from Papay (2011), for a visual representation of this logic. While neither summer misattribution is desirable, the latter (FF) seems less problematic because the teacher has at least had the opportunity to affect the child’s summer following the school year.

Given that the SS versus FF comparison was not the focus of the earlier Papay work, it is unknown whether this finding would be robust to sensitivity checks or would replicate in another setting, in a different subject, or with another assessment. The main focus of the current paper is to explore this very issue. We attempt both to replicate this finding in our data context using the same methods as Papay and to conduct additional analyses to consider whether certain value-added models are less subject to this phenomenon than others. We also consider whether both teacher time-invariant VAM scores and teacher-by-year VAM scores would both be affected.

Relevant Literature

The field is divided as to the utility of teacher value-added estimates in both research and policy contexts: Some research suggests the validity of VAMs may warrant their use (Briggs & Dadey, 2017; Chetty, Friedman, & Rockoff, 2011, 2014a, 2014b; Jacob & Lefgren, 2008; Kane, McCaffrey, Miller, & Staiger, 2013; Koedel & Betts, 2011). Other research has highlighted concerns about such measures (Guarino, Reckase, & Wooldridge, 2011; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Rothstein, 2009). For instance, in his study of teacher VAM scores across different outcome tests, Papay (2011) finds moderate correlations of between 0.15 and 0.58 between estimates derived from three different reading tests in a large urban school district. His study suggests that this variation in estimated teacher effects across outcomes was not explained by factors such as test content, sample of students, item format, or scaling.

As in the current study, Papay (2011) had access to a context in which one of the district's assessments was administered in both fall and spring. Although test timing was not the focus of Papay's paper, he notes that there is an essentially orthogonal relationship ($Corr(FF, SS) = -0.10$) between ELA VAM scores derived from FF and SS timings using the same assessment. He also finds that neither the FF or SS test timings are particularly strongly correlated with what would be the most appropriate—but also typically most infeasible—scenario of using an FS timeline ($Corr(FF, FS) = 0.19$, $Corr(SS, FS) = 0.66$).

There is some precedent for this finding. In an analysis of the effectiveness of compensatory education programs, David and Pelavin (1977, 1978) compare achievement gains for several programs based on traditional FS test timings, which comprise a single school year, and a 12-month FF time period, which includes the summer following the program. The authors also find that estimates of program effectiveness are very sensitive to this timing change and that the inclusion of summer months can substantially lower achievement estimates, sometimes even reversing positive judgements of program effectiveness.

We are aware of one other study that has considered the role of test timing in the estimation of teacher VAM scores. Using the Early Childhood Longitudinal Study-Kindergarten Cohort

(ECLS-K:1999) data, Gershenson and Hayes (2018) examine how VAMs conflate summer and school-year learning. This dataset contains student achievement scores in the fall and spring of grades K and 1 for about one-third of the initial sample. One major constraint of using ECLS:K is that the authors cannot estimate the troubling $Corr(FF, SS)$ which is at the heart of the current paper, and which Papay finds to be essentially zero (-0.10). The authors *are* able to estimate—in kindergarten—the $Corr(FF, FS)$ at 0.80 (ELA) and 0.45 (math). This appears much stronger than Papay's estimate of this same correlation at 0.19 (ELA only). Gershenson and Hayes also estimate—in first grade—the $Corr(SS, FS)$ to be 0.92 (ELA) and 0.80 (math). Again, these are substantively higher than Papay's estimate of the $Corr(SS, FS)$ at 0.66 (ELA). It is unclear why Gershenson and Hayes' findings differ from that of Papay (and the current study). One difference is that, in ECLS-K:1999, each teacher is only observed in one year, whereas in the current analyses, teachers' effects are estimated across up to four years. But perhaps the largest difference between these studies is the grade range: In practice, VAM scores are generated and used almost exclusively for teachers in grades four and above. Given that ECLS-K:1999 can only be used to tackle this question in grades K and 1, Gershenson and Hayes' findings may not generalize to the policy and research contexts in which debates about VAMs take place.

Data and Analytic Sample

The current study uses 2011-12 through 2014-15 administrative student- and school-level data from one anonymous district located in the southeastern US. The district covers one of the state's largest cities and its surrounding area, and it consists of 60 schools serving about 50,000 students annually. It spans a large geographic area—about 245 square miles—which includes a central urban area surrounded by both suburban and rural communities.

The district provides typical demographic data for its students, including student gender, race/ethnicity, Free/Reduced-Price Lunch Program (FRPL) eligibility, Limited English Proficiency (LEP) status, and Special Education (IEP) status. In addition, the district provides a roster that links students to their classrooms, teachers, and schools in each grade and school year. These roster links allow us to construct teacher VAM scores, as described below. Unfortunately, additional teacher covariates, such as years of experience, are not included in the dataset. The district also provides all student-level annual test scores from assessments administered district- or state-wide. A key feature of the current data is that this district administers the Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) assessment *twice* annually to students in grades 3 through 8, in both fall and spring. It is important to acknowledge that the MAP assessment has different purposes and scaling properties than the state's standardized achievement test: MAP is intended to be used as a supplementary tool to aid schools in improving their instruction and meeting students' needs, not as the high-stakes test of record. Because the MAP assessment is designed to monitor students' progress throughout the school year, it is administered in both the fall and the spring (and in some cases also the winter). MAP test results are scored in a manner designed to follow a vertical and interval scale, to produce what NWEA calls a RIT score; in our primary analyses, however, we standardize these scores within subject-grade-semester. A potential limitation of the current study is that we do not make use of the state's high-stakes assessment,[3] which would typically be used to estimate teacher VAM scores.

[3] In this case, the name of the state's standardized test is withheld in order to maintain anonymity. It is administered in the spring of each year from grade 3 through grade 8, and is a test of minimum expectations for English, mathematics, history, and other subjects.

However, in this district there is a clear correspondence between students' standardized MAP and statewide test scores (e.g., in math a correlation of 0.78).

We limit the current analytic sample to students observed with MAP test scores in grades 3 through 8 and their linked math and ELA teachers. To provide context for the kinds of districts to which our findings might generalize, Table 1 contains basic student- and school-level demographics for the analytic sample in one example school year (2011-12).

[Insert Table 1 about here]

This is a diverse analytic sample of about 22,000 third through eighth grade students annually, of whom about half are non-White and over one-third are eligible for the federal Free/Reduced-Price Lunch Program (FRPL). Black students comprise the largest non-White racial/ethnic group (38 percent), while another 8 percent are Asian, and 5 percent are Hispanic. About half (52 percent) of the students are male. About 36 percent are FRPL-eligible, 13 percent have an IEP, and 6 percent are designated as LEP. We also present student, teacher, and school sample sizes across grades and years in Table 2.

[Insert Table 2 about here]

Analytic Approach

We begin by replicating the relevant analyses from Papay (2011) as closely as possible. To do so, we adopt the model shown in his Equation (1) to estimate time-invariant teacher VAM scores, separately for math and English language arts (ELA):

$$Y_{ijt} = \alpha_g f\left(Y_{i,t-1}\right) + \boldsymbol{X'_{it}\gamma} + \boldsymbol{\overline{X}'_{jt}\varsigma} + \delta_j + \varphi_k + \theta_g + \mu_t + \varepsilon_{ijt} \tag{1}$$

Like Papay, we model end-of-period test scores as a polynomial function of a prior test score, student and classroom covariates, as well as teacher (δ_j), school (φ_k), grade (θ_g), and year (μ_t)

fixed effects. We first describe this model using the traditional SS test timing for ELA scores: In Equation (1), the outcome Y_{ijt} represents the ELA MAP score at time $t = end\ of\ spring$ for student i assigned to teacher j in grade g in school k. MAP scores are standardized within subject, grade-semester, and year. On the right-hand side, this outcome is modeled as an up to fifth-order polynomial function of $Y_{i,t-1}$, which is the student's baseline MAP ELA score at time $(t-1) = prior\ spring$. Baseline test score functions are permitted to vary by grade level. The vector $\boldsymbol{X}'_{it}$ of student characteristics includes race/ethnicity dummies, gender, FRPL status, LEP status, and IEP status.[4] We aggregate these same covariates in vector $\overline{\boldsymbol{X}}'_{jt}$ to the teacher-year level. The teacher fixed effects themselves, δ_j, become the teacher VAM scores. We follow Papay's procedure to then convert these into teacher *rankings*, δ'_j, much as a district might do to facilitate comparisons across its teachers.

We estimate the model above three times on the same sample of teachers, but we change how we define $t = end\ of\ period$ and $(t-1) = prior\ period$ baseline score timing (i.e., SS, FF, and FS). We generate for each teacher three rankings: $\delta_j'^{FS}$ (an ideal but impractical option), $\delta_j'^{SS}$ (the only option typically available to districts), and $\delta_j'^{FF}$ (a theoretically preferable alternative to SS). We then produce teacher-level Spearman rank correlations across the three test administration timings; that is, we estimate $Corr(\ \delta_j'^{SS}, \delta_j'^{FF}\ )$, $Corr\big(\ \delta_j'^{FF}, \delta_j'^{FS}\ \big)$, and $Corr(\ \delta_j'^{SS}, \delta_j'^{FS}\ )$. High correlations would suggest that test timing has little influence on which teachers are considered high- versus low-performing and should not be a source of much concern. Correlations near zero indicate that test-timing—a mere artifact of how statewide standardized

[4] Our vector of student demographics is identical to that of Papay, except that he also has access to Gifted and Talented Education (GATE) status, which we do not.

assessments are typically administered—is a highly influential factor in how teachers are ranked according to VAMs.

We extend the analysis beyond replication in several ways. First, while Papay (2011) conducts the analysis only for ELA, we can also do so for math. We also examine whether the strength of these correlations differs depending on how the value-added model is specified (e.g., with school covariates in place of school fixed effects). Additionally, since MAP test scores are vertically scaled, we also run the analysis using the original RIT scores to see whether vertical scaling would attend to the summer learning loss patterns in a way that standardized scores cannot. Finally, we re-specify the value-added model to produce teacher-year VAM scores, δ_{jt} (as opposed to teacher time-invariant scores). This allows us to do two things: First, we can examine whether the estimated correlations are similar when teachers' scores are allowed to vary from one year to another. Second, we can see whether those correlations depend on the teachers' primary grade level taught in a given year.

Primary Results

Like Papay, we find that switching to a different test administration timing dramatically alters the ranking of teachers based on the value-added model specified in Equation (1). We find essentially orthogonal associations between teacher VAM scores produced utilizing SS versus FF test administration timings. We show in Table 3 that, where Papay found the $Corr(\ \delta_j'^{SS}\ , \delta_j'^{FF}\ )$ to be -0.10 (ELA), we find similar correlations of -0.02 for math and 0.07 for ELA.

[Insert Table 3 about here]

To illustrate the point in a way that highlights how this could affect the lives of teachers, we represent this finding in Table 4 as a Q-Q transition matrix. For the rows of Table 4, we categorize teachers according to their SS-based VAM score ranking, $\delta_j'^{SS}$, into one of four quartiles (Q1 is highest, Q4 is lowest). We do the same for the columns of Table 4, but instead using their FF-based VAM score ranking, $\delta_j'^{FF}$, to produce quartiles. The cells of Table 4 contain the number of teachers in each unique combination of quartiles, as well as the corresponding row percentages. For instance, 42 of the 134 ELA teachers (31 percent) who are in the top quartile according to $\delta_j'^{SS}$ are also in the top quartile according to $\delta_j'^{FF}$.

[Insert Table 4 about here]

Teachers along the diagonal of the transition matrix appear in the same quartile using either their SS-based or FF-based VAM score rankings. However, only 30 percent of all ELA teachers appear along this diagonal (29 percent for math). For the other 70 percent of teachers, they would have been placed into different quartiles of performance under these two test timings. Further, for 33 percent of ELA teachers, their FF-based VAM score would place them in a quartile two or more quartiles away from their SS-based quartile. In fact, 27 percent of top-quartile ELA teachers according to $\delta_j'^{SS}$ appear in the *bottom* quartile according to $\delta_j'^{FF}$, and 21 percent of bottom-quartile ELA teachers according to $\delta_j'^{SS}$ appear in the *top* quartile according to $\delta_j'^{FF}$. In essence, a sizable portion of the seemingly weakest teachers according to the traditional SS-timing would be categorized among the strongest teachers according to the FF-timing. Results are nearly identical for math. Taken together, we see that FF-based teacher VAM score rankings bear little resemblance to those based on the traditional SS-based timing. Given that the FF-based timing would be theoretically preferable[5] to SS-based timing, which is near-universally the only practical

[5] Recall that the FF timing incorporates into a teacher's "effect" on a student the summer *after* they encounter a given student (which the teacher could at least possibly influence), as opposed to SS timing

option, we should be concerned about how the constraints of statewide testing systems may inadvertently compromise attempts to estimate and compare teacher effects.

We also consider the use of a more ideal FS-based timing for VAM scores, in which end-of-year spring test scores are modeled as a function of start-of-year fall test scores, thus isolating both the summer before and after the school year from the teacher effects. The downside, of course, of the FS-based value-added model is that it would require standardized testing twice per year. We are interested in whether this ideal scenario ($\delta_j'^{FS}$) aligns more with the typically-available option ($\delta_j'^{SS}$), or the theoretically-preferred $\delta_j'^{FF}$ alternative. In Table 3, we find the $Corr\left(\ \delta_j'^{SS}\ , \delta_j'^{FS}\ \right)$ to be 0.31 for ELA and 0.07 for math. For the $Corr\left(\ \delta_j'^{FF}\ , \delta_j'^{FS}\ \right)$, we find somewhat stronger correlations of 0.49 for ELA and 0.57 for math. These patterns again mirror the Papay findings for ELA teachers, wherein the weakest association was for $Corr(\ \delta_j'^{SS}\ , \delta_j'^{FF}\ )$ at -0.10, followed by $Corr(\ \delta_j'^{FF}\ , \delta_j'^{FF}\ )$ at 0.19, and then the strongest association for $Corr(\ \delta_j'^{SS}\ , \delta_j'^{FF}\ )$ at 0.66. The fact that the patterns in the current replication so closely resemble those found by Papay (2011) lends support to the possibility that this finding was not idiosyncratic to that setting.

Robustness Checks

We consider the possibility that the low correspondence of VAM scores across test timings is an artifact of the specific value-added model chosen by Papay and shown in Equation (1). This is particularly salient if that model is unlikely to appear in a real-world context. For instance, Equation (1) includes school fixed effects, and while there are often reasons to include these for

which incorporates the summer *before* the teacher encounters the students into the teacher's effect on that student.

research purposes, districts are unlikely to use such a model since it undermines the ability to compare teachers across schools. In Table 5, we therefore reproduce these three correlations—$Corr(\delta_j'^{SS}, \delta_j'^{FF})$, $Corr(\delta_j'^{FF}, \delta_j'^{FS})$, and $Corr(\delta_j'^{SS}, \delta_j'^{FS})$—across eight different permutations of the value-added model for both ELA (left) and math (right). Note that "Model 5" mirrors Papay's VAM specification, as well as the results presented above in Table 3 and Table 4. We conduct this analysis both using standardized versions of the MAP scores (upper panel), as well as MAP's original, vertically-scaled metric, the RIT score (lower panel).

[Insert Table 5 about here]

Overall, we find that the general takeaway holds across model specifications. Though there is some variability across models, results generally reflect the magnitudes found by Papay. In every model, the $Corr(\delta_j'^{SS}, \delta_j'^{FS})$ is the strongest of the three, the $Corr(\delta_j'^{FF}, \delta_j'^{FS})$ is in the middle, and the $Corr(\delta_j'^{SS}, \delta_j'^{FF})$ is the weakest of the three. In most cases, the $Corr(\delta_j'^{SS}, \delta_j'^{FF})$ is either slightly negative (as found in Papay) or less than 0.15 and thus nearly independent. Across all eight models, two subjects, and both scalings of the scores, the $Corr(\delta_j'^{SS}, \delta_j'^{FF})$ is never stronger than 0.34. Regardless of the value-added model specification, SS test timing will provide a very different ranking of teachers from those based on an FF test timing. However, it does seem to be the case that—among these low correlations—models that include school fixed effects (Models 5 through 8) tend to exhibit the weakest correlations between $\delta_j'^{SS}$ and $, \delta_j'^{FF}$. The strongest estimated correlations appear for $Corr(\delta_j'^{SS}, \delta_j'^{FS})$ in models without school fixed effects, the highest of which is 0.72. However, the mean of all correlations presented in Table 5 is only 0.33.

The analyses above are intended to closely mirror Papay's methods to determine whether his findings with regard to teacher effects would replicate in the current district using a different assessment. It appears they do. However, if one's sole focus is to understand the current phenomenon, one might prefer to estimate teacher-*by-year* VAM scores in order to more directly consider how absorbing the preceding summer instead of the following summer affects how a teacher is evaluated *in a given year*. Hypothetically, we would expect that SS-based VAM scores could be most biased when teachers with very different latent effectiveness inherit one another's students. In a simplified scenario to illustrate the point, imagine that a truly ineffective fourth grade teacher has inherited a classroom of students taught by a truly effective third grade teacher in the previous year. We may hypothesize that the third-grade teacher would have a lasting impact on her students that persists into the summer after third grade—a summer which, under the SS-timing, will be inappropriately attributed to the fourth-grade teacher. The greater the differential in latent effectiveness (which is unobservable to us) between a given teacher and the teacher(s) who taught their students in the preceding year, the more one might expect FF-based, SS-based, and FS-based VAM scores to diverge.

We therefore rerun the model presented in Equation (1) but now substitute teacher-by-year fixed effects, δ_{jt}, for the teacher fixed effects, δ_j. This produces a VAM-based ranking for each teacher in each year, thus allowing a given teacher to vary in terms of estimated effectiveness across the four years of the panel. We reproduce Table 3, but now conduct the correlations at the teacher-year unit of analysis, rather than the teacher level. This also allows us to disaggregate the results by the teacher's modal grade taught in each year. Results are presented in Table 6.

[Insert Table 6 about here]

We continue to find that the correspondence between FF-based and SS-based VAM score rankings tends to be low: The $Corr(\delta_{jt}'^{SS}, \delta_{jt}'^{FF})$ for teachers across all grades is 0.10 for ELA and -0.02 for math. When turning to the grade-specific results, there do appear to be substantive differences in the $Corr(\ \delta_{jt}^{\prime SS}\ ,\delta_{jt}^{\prime FF}\ )$ across grades: In ELA, while the correlation is 0.02 for grade 7 teachers, that same correlation is 0.35 for grade 8 teachers. However in math, the $Corr(\ \delta_{jt}^{\prime SS}\ ,\delta_{jt}^{\prime FF}\ )$ is near zero in grades 5, 7, and 8, but a surprising -0.43 in grade 6. In general, there does not appear to be a clear trend across grades in the strength of these associations, and we see less correspondence in these results across subjects than in preceding tables. The disaggregation by grade shows some surprising results that would be worthy of more in-depth investigation than is possible with the current data. At the same time, the results in Table 6 generally continue to corroborate the main finding that measures of teaching effectiveness strongly depend on whether the underlying value-added model utilizes a FF, SS, or FS test administration timing.

Discussion

We find that there is virtually no relationship between how teachers are ranked based on VAM scores produced by an otherwise identical value-added model that estimates student growth *either* from prior-spring to current-spring ("SS"—the pragmatically feasible option) or current-fall to next-fall ("FF"—a theoretically better alternative to SS). Moreover, neither of these options exhibits a strong correlation with VAM score rankings based on a current-fall to current-spring ("FS"—ideal but impractical) test timing. Results from the current study appear to replicate findings from the similar analysis embedded in Papay (2011). Moreover, the findings generally hold up against the additional robustness checks we are able to bring to bear.

These findings have clear policy implications for how VAM scores are used to characterize teacher performance in both policy and research contexts. Value-added measures of teacher effectiveness often play a large role in high-stakes decisions about teacher retention, placement, and compensation. Low correlations between models that—at least intuitively—should produce somewhat similar rankings calls into question the fairness of using SS-based VAM scores in ways that can dramatically impact the lives of teachers. For example, districts that use value-added measures as a factor in teacher compensation may not be correctly identifying the teachers who have the strongest positive impact on student achievement outcomes.

A clear candidate explanation for the disparities in VAM scores based on FF- versus SS-based test timings lies in the potential misattribution of summer periods. Several studies have shown substantial decline in student achievement scores during summer vacation; furthermore, this decline appears to disproportionately occur for low-income and historically marginalized students (Allington et al., 2010; Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; Entwisle & Alexander, 1992). In preliminary results not shown (available upon request), we find some evidence that teachers working in high-FRPL schools may be differentially affected by these timing-based VAM discrepancies. A fruitful next step would be to explore whether disaggregating across certain kinds of teachers, schools, or grade configurations helps reveal underlying patterns in these correlations that elucidate why they are so low.

The current findings contribute to the existing literature that considers the many factors that drive the same teacher to have somewhat different value-added scores (e.g., year-to-year instability, in math versus ELA, different test outcomes, controlling for various sets of observed covariates, including student or school fixed effects, using a random effects framework, or a student growth percentile approach, etc.) In general terms, the literature documents concerningly low correlations in the approximate range of 0.20 to 0.70. Overall, at least, VAM scores for the same teacher do tend to move in similar *directions*. In contrast, the *near-zero* correlations focused upon in the current paper raise concerns of a different magnitude. If these findings truly reflect an unprincipled artifact of test administration timing, then it would suggest that any inferences about teacher effectiveness made based on SS-based to date should be called into question.

However, there are limitations to the current analysis. One key concern is whether or not students and teachers take the fall administration of the MAP assessment as seriously as the spring administration. If fall scores are simply unreliable or inaccurate measures of students' math or ELA skills at that point in time, then incorporating them into a value-added model would not be advisable. In the current context, we cannot fully address this limitation, though we do observe that fall MAP scores are strongly correlated with students' high-stakes test scores on the statewide assessment just a few months earlier in the preceding spring (e.g., 0.79 for math). In future work, we hope others can address this limitation by replicating the analyses in settings where either there are compelling reasons to believe the fall and spring test administrations are of equal importance to teachers, or where information about the reliability of individual test scores—e.g., test score standard errors, seconds spent per item, total minutes spent by each student—are available to the researcher.

The finding of a $Corr(\delta_j'^{SS}, \delta_j'^{FF})$ near zero has now been documented in both a large, urban school district in the U.S. Northeast (Papay) as well as a Southeastern school district of similar size that includes urban, suburban, and rural areas. Future studies can explore whether this important finding can be replicated in other contexts and with other assessments. If so, then researchers and policy-makers need to carefully consider whether it is prudent to use SS-based VAM scores to characterize teachers' relative effectiveness. This is particularly troubling given that the theoretically more appropriate approach of estimating teacher VAM scores based on FS or FF timings is simply not an option for most districts.

Tables and Figures

Table 1. Sample Descriptives at Student- and School-Level, AY 2011-12

	Mean	SD	N
Student-Level Characteristics			
% Male	0.52	--	23,308
% Asian	0.08	--	22,438
% Black	0.38	--	22,438
% Hispanic	0.07	--	22,438
% White	0.47	--	22,438
% LEP	0.06	--	22,438
% FRPL	0.36	--	22,438
% Special Ed.	0.13	--	22,438
Average School-Level Characteristics			
Avg.% Male	0.50	0.07	57
Avg.% Asian	0.08	0.08	56
Avg.% Black	0.38	0.32	56
Avg.% Hispanic	0.07	0.06	56
Avg.% White	0.46	0.28	56
Avg.% LEP	0.07	0.08	56
Avg.% FRPL	0.38	0.25	56
Avg.% Special Ed.	0.13	0.05	56
Number of Students	409	307	57

Table 2. Student, Teacher, and School Sample Sizes, by Year

Students	2011-12	2012-13	2013-14	2014-15	Row Total
Grade 3	3,850	3,929	3,827	4,069	15,675
Grade 4	3,880	3,842	3,876	3,846	15,444
Grade 5	3,844	3,846	3,765	3,908	15,363
Grade 6	3,969	3,929	3,955	3,959	15,812
Grade 7	3,870	3,964	3,963	4,041	15,838
Grade 8	3,922	3,888	4,055	3,991	15,856
Total	*23,335*	*23,398*	*23,441*	*23,814*	*93,988*
Teachers					
ELA	691	691	672	669	2723
Math	695	694	675	662	2726
Total Unique	*829*	*829*	*810*	*800*	*3268*
Schools					
Total	*57*	*56*	*56*	*56*	*225*

Table 3. Teacher-Level Spearman Rank Correlations across FF, SS, and FS Timings

	Correlations			Teacher Sample Sizes		
	Papay Study	Replication	Replication	Papay Study	Replication	Replication
Comparison	(ELA)	(ELA)	(Math)	(ELA)	(ELA)	(Math)
Corr(δ'^{SS}_j , δ'^{FF}_j)	–0.10	0.07	-0.02	624	497	496
Corr(δ'^{FF}_j , δ'^{FS}_j)	0.19	0.31	0.07	713	701	706
Corr(δ'^{SS}_j , δ'^{FS}_j)	0.66	0.49	0.57	684	685	676

FN. Deltas, δ_j', represent the VAM score rankings of teachers (j). These VAM scores are derived from Equation (1) value-added model, which includes student- and classroom controls, as well as teacher, grade, year, and school fixed effects. For the reader's reference, we also include the correlations from the Papay (2011) study, alongside our replication of those correlations.
SS= prior spring → current spring; FF= current fall → next fall; FS=current fall → current spring.

Table 4. Teacher Transition Matrix across Quartiles of SS vs. FF Test Timings, by Subject

ELA

	(Top)			(Bottom)	Row
	FF - Q1	FF - Q2	FF - Q3	FF - Q4	Total
SS - Q1 (Top)	42	28	28	36	134
	(0.31)	*(0.21)*	*(0.21)*	*(0.27)*	*(1.00)*
SS - Q2	28	45	32	24	129
	(0.22)	*(0.35)*	*(0.25)*	*(0.19)*	*(1.00)*
SS - Q3	22	36	33	30	121
	(0.18)	*(0.30)*	*(0.27)*	*(0.25)*	*(1.00)*
SS - Q4 (Bottom)	24	28	32	29	113
	(0.21)	*(0.25)*	*(0.28)*	*(0.26)*	*(1.00)*
Col Total	116	137	125	119	497
	(0.23)	*(0.28)*	*(0.25)*	*(0.24)*	*(1.00)*

MATH

	(Top)			(Bottom)	Row
	FF - Q1	FF - Q2	FF - Q3	FF - Q4	Total
SS - Q1 (Top)	34	34	24	33	125
	(0.27)	*(0.27)*	*(0.19)*	*(0.26)*	*(1.00)*
SS - Q2	25	35	43	26	129
	(0.19)	*(0.27)*	*(0.33)*	*(0.20)*	*(1.00)*
SS - Q3	21	34	48	28	131
	(0.16)	*(0.26)*	*(0.37)*	*(0.21)*	*(1.00)*
SS - Q4 (Bottom)	34	30	20	27	111
	(0.31)	*(0.27)*	*(0.18)*	*(0.24)*	*(1.00)*
Col Total	114	133	135	114	496
	(0.23)	*(0.27)*	*(0.27)*	*(0.23)*	*(1.00)*

FN: Quartiles are created based on the model used in Table 3 and the value-added model specified in Equation (1). Cells contain counts of teachers in a given quartile transition (e.g., for ELA, 36 teachers were in Q1 according to SS timing but Q4 according to FF timing). Row percentages are included in parentheses in cells. We also present row and column totals at the margins.
SS= prior spring → current spring; FF= current fall → next fall.

Table 5 Robustness Check: Spearman Rank Correlations for Same Model and Different Timings, for 8 Different VAM Specifications

	Controls in VA Model				ELA Correlations			Math Correlations		
	Student	Class	School	Sch Effects	$(\delta'^{SS}_j, \delta'^{FF}_j)$	$(\delta'^{FF}_j, \delta'^{FS}_j)$	$(\delta'^{SS}_j, \delta'^{FS}_j)$	$(\delta'^{SS}_j, \delta'^{FF}_j)$	$(\delta'^{FF}_j, \delta'^{FS}_j)$	$(\delta'^{SS}_j, \delta'^{FS}_j)$
Papay Findings	X	X		X	–0.10	0.19	0.66	--	--	--
Standardized Scores										
Model 1	X	X	X		0.11	0.20	0.68	0.20	0.30	0.68
Model 2		X	X		0.13	0.22	0.69	0.20	0.29	0.66
Model 3	X		X		0.16	0.22	0.70	0.32	0.34	0.70
Model 4			X		0.21	0.28	0.72	0.34	0.36	0.69
Model 5	X	X		X	0.07	0.31	0.49	-0.02	0.07	0.57
Model 6		X		X	0.03	0.34	0.47	-0.01	0.08	0.55
Model 7	X			X	0.14	0.34	0.53	0.08	0.16	0.60
Model 8				X	0.20	0.41	0.54	0.16	0.24	0.60
RIT Scale Scores										
Model 1	X	X	X		0.06	0.20	0.67	0.21	0.33	0.69
Model 2		X	X		0.07	0.22	0.67	0.20	0.33	0.66
Model 3	X		X		0.11	0.24	0.69	0.32	0.37	0.71
Model 4			X		0.17	0.30	0.71	0.32	0.38	0.70
Model 5	X	X		X	0.04	0.29	0.46	-0.03	0.04	0.54
Model 6		X		X	-0.01	0.30	0.44	-0.04	0.05	0.53
Model 7	X			X	0.11	0.34	0.49	0.06	0.15	0.57
Model 8				X	0.17	0.42	0.50	0.13	0.24	0.58

FN: The findings from Papay (2011) are shown in the first row (note that the Papay VA model is akin to "Model 5" in this table). We iterate through 8 permutations of the VA model. "Student" indicates that the model included student control (race/ethnicity dummies, gender, FRPL status, LEP status, and IEP status). "Class" indicates the model include those same covariates aggregated to the classroom level. "School" indicates the inclusion of that the same vector of student-level covariates aggregated to the school level, as well as school size. "Sch Effects" indicates that school fixed effects were included in lieu of school-level controls.
SS= prior spring → current spring; FF= current fall → next fall; FS= current fall → current spring.

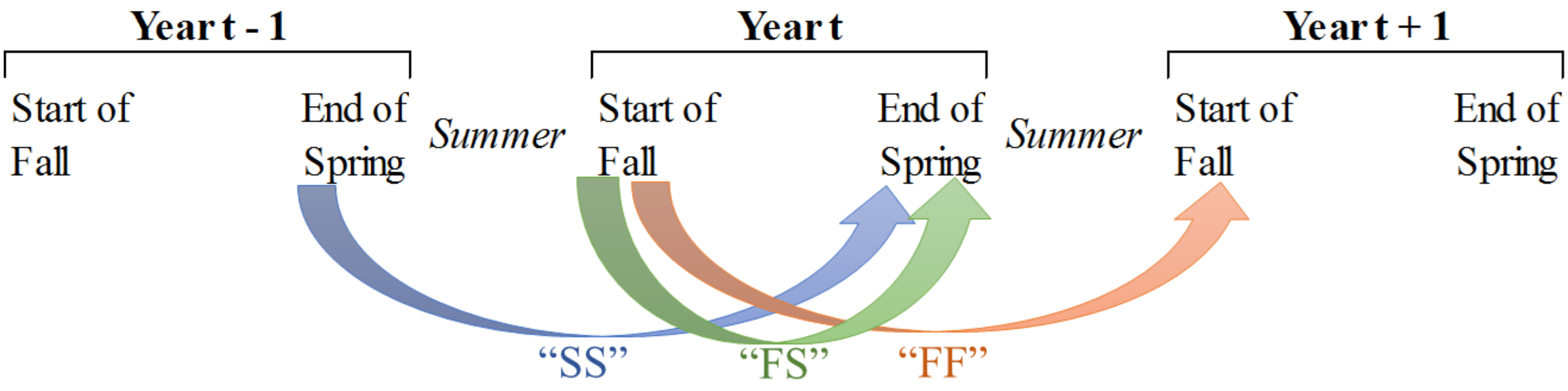
Table 6. Teacher-by-Year VAM Score Ranking Correlations, Overall and by Grade.

	Papay	Teacher's	Correlations	
	Study	Grade	ELA	Math
Corr(δ'^{SS}_{jt}, δ'^{FF}_{jt})	–0.10	All Grades	0.10	-0.02
		5	0.07	0.05
		6	0.24	-0.43
		7	0.02	0.02
		8	0.35	-0.02
Corr(δ'^{FF}_{jt}, δ'^{FS}_{jt})	0.19	All Grades	0.23	-0.01
		5	0.26	-0.09
		6	0.36	-0.04
		7	0.34	0.29
		8	0.31	0.26
Corr(δ'^{SS}_{jt}, δ'^{FS}_{jt})	0.66	All Grades	0.55	0.58
		5	0.41	0.44
		6	0.77	0.60
		7	0.76	0.63
		8	0.66	0.77

FN. Deltas, δ'_{jt}, represent the VAM score rankings of teacher (j) in year (t). These VAM scores are derived from a VA model which includes student- and classroom controls, as well as grade, year, and school fixed effects. Importantly, the teacher fixed effects from Equation (1) are replaced with teacher-by-year fixed effects. Sample sizes for grade-specific correlation presented in the table range from about 100 to 300. For the reader's reference, we also include the original teacher-level correlations from the Papay (2011) study.

SS= prior spring → current spring; FF= current fall → next fall; FS=current fall → current spring.

Figure 1. Illustration of 3 Possible Combinations of Pretest and Outcome Tests to Estimate Teacher's Effect in Year t



FN: This is a slightly modified version of Figure 1 from Papay (2011), reproduced here with permission.

References

Allington, R. L., McGill-Franzen, A., Camilli, G., Williams, L., Graff, J., Zeig, J., . . . Nowak, R. (2010). Addressing Summer Reading Setback Among Economically Disadvantaged Elementary Students. *Reading Psychology, 31*(5), 411-427. doi:10.1080/02702711.2010.505165

Briggs, D. C., & Dadey, N. (2017). Principal holistic judgments and high-stakes evaluations of teachers. *Educational Assessment, Evaluation and Accountability, 29*(2), 155-178.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. National Bureau of Economic Research.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review, 104*(9), 2593-2632. doi:doi: 10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review, 104*(9), 2633-2679. doi:doi: 10.1257/aer.104.9.2633

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.

David, J. L., & Pelavin, S. H. (1977). *Research on the effectiveness of compensatory education programs: A reanalysis of data*. Retrieved from ERIC:

David, J. L., & Pelavin, S. H. (1978). Evaluating Compensatory Education: Over What Period of Time Should Achievement Be Measured? *Journal of Educational Measurement, 15*(2), 91-99.

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*, 72-84.

Gershenson, S., & Hayes, M. S. (2018). The Implications of Summer Learning Loss for Value-Added Estimates of Teacher Effectiveness. *Educational Policy*, 31.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2011). *Evaluating Value-Added Methods for Estimating Teacher Effects*. Paper presented at the Society for Research on Educational Effectiveness, Washington, DC.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

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

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

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

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537-571.