education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 23 Number 92 September 28th, 2015 ISSN 1068-2341

The Effect of Summer on Value-added Assessments of Teacher and School Performance

Gregory J. Palardy
&
Luyao Peng
University of California, Riverside
United States

Citation: Palardy, G. J., & Peng, L. (2015). The effects of including summer on value-added assessments of teachers and schools. *Education Policy Analysis Archives, 23*(92). http://dx.doi.org/10.14507/epaa.v23.1997

Abstract: This study examines the effects of including the summer period on value-added assessments (VAA) of teacher and school performance at the early grades. The results indicate that 40-62% of the variance in VAA estimates originates from the summer period, depending on the outcome (i.e., reading or math achievement gains). Furthermore, when summer is omitted from the VAA model, 51-61% of the teachers and 58-61% of the schools change performance quintiles, with many changing 2-3 quintiles. Extensive statistical controls for student background and classroom and school context reduce the summer effect, but 36-47% of the teachers and 42-49% of the schools are still in different quintiles. Furthermore, besides misclassifying teachers and schools, the results show that including summer tends to bias VAA estimates against schools with concentrated poverty. The results suggest that removing summer effects from VAA estimates will likely require biannual achievement assessments (i.e., fall and spring).
Keywords: accountability; school effectiveness; teacher evaluation; high stakes testing.

Los efectos de la inclusión del verano en las evaluaciones de valor añadido de docentes y escuelas.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 2/4/2015
Revisions received: 6/10/2015
Accepted: 6/10/2015

Resumen: Este estudio analiza los efectos de incluir el período de verano en las evaluaciones de valor añadido (EVA) de docentes y de rendimiento escolar en los primeros grados. Los resultados indican que el 40-62% de la varianza en las estimaciones de EVA se originan en el período de verano, dependiendo del resultado (es decir, ganancias de logros en lectura o en matemáticas). Por otra parte, cuando el verano se omite del modelo EVA, desde 51-61% de los profesores y de 58-61% de las escuelas cambian sus quintiles de desempeño, y muchas cambian 2-3 quintiles. Controles estadísticos extensivos de las características de los estudiantes y de contexto escolar reducen el efecto de verano, pero 36-47% de los profesores y de 42-49% de las escuelas están todavía en diferentes quintiles. Estos resultados sugieren que la eliminación de los efectos de verano, de las estimaciones EVA probablemente requerirá evaluaciones de rendimiento bianuales (es decir, de otoño y primavera). Por otra parte, además de clasificar erróneamente a docentes y las escuelas, los resultados muestran que la inclusión de verano tiende a sesgar las estimaciones EVA en contra de las escuelas con mayores concentraciones de estudiantes en situación de pobreza.
Palabras clave: rendición de cuentas; eficacia de la escuela; evaluación docente; exámenes de consecuencias severas.

Os efeitos da inclusão de verão nas avaliações de valor agregado de professores e escolas.
Resumo: O presente estudo analisa os efeitos de incluir o período de Verão nas avaliações de valor agregado (AVA) de professores e desempenho escolar nas séries iniciais. Os resultados indicam que 40-62% da variação nas estimativas de AVA se originam no período de verão, dependendo do resultado (ou seja, os ganhos de desempenho em leitura ou matemática). Além disso, quando o verão é omitido do modelo de AVA, 51-61% dos professores e 58-61% das escolas mudam um quintil de desempenho, e muitos mudaram 2-3 quintis. Verificações estatísticas abrangentes sobre características dos estudantes e do contexto escolar reduzem o efeito do verão, mas 36-47% dos professores e 42-49% das escolas ainda estão em quintis diferentes. Estes resultados sugerem que a eliminação dos efeitos de verão em AVA provavelmente exigira avaliações de desempenho bianuais (isto é, outono e primavera). Além disso, além de classificar erroneamente professores e as escolas, os resultados mostram que a inclusão de verão tende a distorcer as estimativas AVA contra as escolas com maiores concentrações de estudantes em situação de pobreza.
Palavras-chave: prestação de contas; a eficácia da escola; avaliação de professores; testes de consequências graves.

Introduction

As the educational accountability movement gained traction over the past three decades, federal, state, and local policies have increasingly tied teacher and school performance assessment to student achievement test scores. These policies have led to considerable research and debate on how to best gauge the contributions that individual teachers and schools make to their students' achievement. In recent years value-added assessment (VAA) has emerged as the most recommended statistical approach for this purpose (Glazerman et al., 2010; Harris, 2011; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2003; Tekwe et al., 2004). Consequently, the application of VAA for teacher and school accountability assessment has experienced an enormous expansion over the past decade.

The increased use of VAA has not gone without criticism, particularly for high stakes applications that may result in sanctions against "low performing" teachers or schools. The primary concerns are that VAA estimates can be unreliable and that measurement issues with achievement test scores can bias VAA estimates. For instance, research shows that VAA estimates of teacher

effectiveness tend to vary substantially from year to year and from one standardized achievement test to another (Lockwood et al., 2007; Papay, 2011). Moreover, measurement issues with achievement tests, such as ceiling or floor effects, nonlinearity in the test's scale, or imperfections in the vertical equating of tests that are used to estimate change in achievement over time, can bias VAA estimates (Haertel, 2013; Koedel & Betts, 2010; Reardon & Raudenbush, 2009). These and other shortcomings have led several prominent education scholars to advise against using VAA for high-stakes personnel decisions (Amrein-Beardsley, 2008; Baker et al., 2010; Braun, Chudowsky, & Koenig, 2010; McCaffrey et al., 2003).

One factor that potentially impacts VAA that has not received adequate research attention is the inclusion of the summer period when students are not attending school. This is a noteworthy deficit in the literature because VAA is typically based on annual gains in student achievement from one spring to the next, which includes the summer period when teachers and schools tend to have little control over what students do. This is problematic because research indicates that student achievement tends to drop over summer and that demographic achievement gaps primarily develop during summer (Alexander, Entwisle, & Olson, 2001; Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; Heyns, 1978). Hence, using annual spring-to-spring assessments may introduce variation originating from summer that biases VAA estimates. Moreover, given that the rate at which achievement gaps develop accelerates over summer, VAA estimates that include summer may be biased against teachers and schools serving inordinately high proportions of disadvantaged children (Baker et al., 2010; Harris, 2011). However, the degree to which including summer impacts VAA estimates and whether estimates are biased against teachers and schools serving disadvantaged populations remains unclear. It is also unclear whether any summer effects on VAA can be ameliorated using statistical control covariates that are typically available for accountability modeling (e.g., student demographics and free or reduced lunch status).

Research Questions

The present study examines the impact of including the summer period on VAA estimates of teacher and school performance based on gains in student reading and math achievement test scores. VAA estimates derived from the typical annual achievement gains testing schedule (spring of one year to spring of the next year) are compared with VAA estimates derived from the school year gains testing schedule (fall to spring of the same school year). Comparisons are based on correlations, the proportion of the variance in VAA estimates derived from annual-gains that originates from summer, and quintile classification differences. A nationally representative sample of first graders and their teachers and schools were used to address the following research questions:

1. To what extent does including summer impact VAA estimates of teacher and school performance?
2. Can any summer effect be ameliorated without biannual assessments (i.e., fall and spring) using control covariates that are typically available to school districts, such as student demographics and contextual characteristics of classrooms and schools?
3. To what degree does including summer in VAA estimates result in biases against teachers and schools serving low income and ethnic minority children?

Background

Summer Learning and Achievement Gaps

Research has documented substantial differences in the rate at which children learn during the school year compared with over summer when they are not attending school (Alexander, Entwisle, & Olson, 2001; Borman & Boulay, 2004; Cooper et al., 1996; Heyns, 1978). In one of the first major studies on this topic, Heyns (1978) conceptualized student achievement as the result of innate ability and a mixture of three environmental influences: home, community, and school. Whereas the home and community factors are essentially year-round influences, the effect of schooling is mostly limited to when school is in session. Heyns found that the socioeconomic achievement gap primarily develops over summer, suggesting it is largely the product of socioeconomic differences in home and community influences. She concluded that because children spend far less time in those settings when school is in session, the rate of increase in socioeconomic achievement gaps tends to slow dramatically during the school year.

Over the past 30 years several additional studies have been conducted on summer effects that mostly support Heyns' findings (Alexander et al., 2001; Borman & Boulay, 2004; Cooper et al., 1996). A meta-analysis by Cooper et al. (1996) of 13 select studies found that, on average, students lose approximately one grade-equivalent month of achievement over summer, although the magnitude of the summer loss tends to be larger for math than reading. Moreover, the meta-analysis indicated that the summer loss is associated with SES, but only on reading. In fact, middle-class students tended to show summer gains in reading achievement. However, some recent research suggests that socioeconomic and ethnic achievement gaps do increase during the school year and not just over summer (Palardy, 2015; Palardy & Rumberger, 2008) and that year-school increases are due in part to differences in instructional practices and teacher effectiveness (Betts, Zau, & Rice, 2003; Murnane, Willett, Bub, & McCartney, 2006; Stipek, 2004). If, as these recent studies conclude, achievement gaps increase during the school year and teachers contribute to that increase, it is less clear whether VAA that include summer in will bias estimates against teachers and schools serving children from educationally disadvantaged backgrounds. It is also unclear as to whether any such bias in VAA estimates can be addressed by controlling for student demographics.

Value-Added Models and Summer Bias

An implication of the research on summer learning to the assessment of school and teacher performance is that the timing of achievement test administration impacts estimates of achievement gains and, by extension, may impact VAA estimates. Consistent with this implication, a recent study concluded that test timing is the largest single source of measurement error and instability in VAA of teacher effectiveness; it is more important than the specification of the model, the sample of students, or the achievement test used (Papay, 2011). One may assume that the optimal testing schedule for VAA will provide data to accurately estimate change in achievement during the period school is in session. Because American schools are typically not in session for several weeks over summer, optimal estimation of VAA may require a minimum of two annual achievement tests, one administered in fall near the beginning of the school year and one in spring near the end. However, fall achievement testing is uncommon in U.S. schools, which has resulted in VAA typically being based on a spring-spring testing schedule.

Considering the literature on summer learning and its potential implications to VAA, there is surprisingly little research on the impact of including summer on VAA estimates. However, one study found that including summer can impact school VAA quintile rankings. Downey, von Hippel, and Hughes (2008) found that of schools classified in the bottom or top VAA quintile when

summer was included, 20% to 35%, depending on the achievement test subject, were classified in another quintile when summer was excluded. This suggests that a substantial percentage of the schools that are classified as failures or successes based on a spring-spring achievement gains are classified as satisfactory when assessed based on a fall-spring achievement gains. While their results were highly revealing, Downey et al. (2008) limited their focus to schools, omitting teacher performance assessment, and did not investigate whether summer effects on school performance can be reduced using control covariates or whether including summer results in biased VAA estimates against high-poverty schools. Another recent study on VAA that speaks to the issue of test timing, argued that ignoring the summer period in VAA is tantamount to ignoring non-linearity in a growth model (Palardy, 2010). The results of that study indicated that ignoring non-linearity in VAA will inflate the variance in teacher effectiveness and bias VAA estimates against teachers and schools whose students have the most negative summer achievement gains (Palardy, 2010). Given the prevalent use of VAA in the U.S., more research is needed to better understand the effects of including the summer period.

Methodology

Data Source

This study uses data from ECLS-K, a nationally representative and longitudinal sample of 1998 kindergarteners, their parents, teachers, and schools (NCES, 2002).[1] Several characteristics of ECLS-K make it highly suitable for addressing the research questions of this study. First, the student sample is approximately nationally representative. This is desirable because accountability practices are commonly implemented in response to federal legislation (e.g., No Child Left Behind). Having a national sample, as opposed to a local sample, broadens the generalizability of the results so that they are more applicable to federal policy. Second, ECLS is the only national database that includes both fall and spring student achievement test scores, which is necessary for studying summer effects. Third, these test scores were set to an interval scale for each testing period and vertically scaled using item response theory (IRT) methods across testing periods. Interval scaled test scores are essential for assuring the gain unit is equivalent across the distribution or scores, while vertical scaling links tests of different difficulty such as the kindergarten and first grade tests. Fourth, ECLS-K includes many measures of student demographics and classroom and school context that are necessary for examining the viability of using control covariates to address any summer effects on VAA estimates.

The ECLS-K first grade longitudinal sample has 5,034 children. Students without teacher or school IDs were omitted, as were a small number of students who had missing test scores or who changed schools during first grade. We also limited our analysis to public schools because federal accountability legislation typically applies to the public sector and because private schools are more prone to selectivity biases that can confound VAA estimates.[2] The sample for the present study included 2,251 students, 682 classrooms, and 168 schools.[3]

---

[1] For more information on ECLS, please see http://nces.ed.gov/ecls/kindergarten.asp

[2] To investigate whether these selection criteria biased our sample, the weighted full first grade longitudinal sample and the weighted sample used in the present study were compared on key variables, including achievement outcomes, SES, and proportion black and Hispanic. Variable means (or proportions) and standard deviations were highly similar in the two samples; no statistical differences were found.

[3] This study uses the weight for the longitudinal first grade sample (C3C4cw0) so that the student sample is approximately nationally representative. However, models were also run without weights. A comparison showed that the weighted and unweighted VAA estimates differed only to a minute degree.

Value-added Assessment Models

All VAA models used in this study have the same general form, differing only in terms of the covariates that are included. The general form was selected based on its strong performance for recovering value-added estimates in two recent simulation studies (Guarino, Reckase, & Wooldridge, 2015; Henry, Rose, & Lauen, 2014). It is a three-level hierarchical linear model (HLM) with an outcome of year-over-year (YoY) or school-year (SY) achievement gains in reading or math. Levels one, two, and three correspond to students, classrooms, and schools, respectively. An advantage of the three-level HLM, as opposed to a two-level HLM, is that teachers are effectively compared with other teachers working in the same school. This helps separate teacher effects from school effects. The teacher and school VAA estimates are the level-two and level-three residuals, respectively. The teacher residuals are essentially the mean gains of the students in the respective teacher's classroom adjusted for the covariates that are included in the model. Similarly, the school residuals are essentially the mean classroom gains at the respective school, again adjusted for the covariates in the model.[4] Henry et al. (2014) used a three-level HLM that is highly similar to that of the present study, which they found to perform better than five other commonly used and highly sophisticated models they tested for recovering teacher value-added estimates. Details on the model, model building, and model specification are provided next.

Model building. For each outcome, four sequential models were estimated: null, base, demographics, and context. Each subsequent model includes the covariates from the previous model plus a new set of covariates. The null model only includes a covariate that adjusts for student differences in the amount of time between the first and second achievement test administration, which varied across schools. Compared with the null model, the base model has only one additional covariate: a measure of achievement at the start of the gain score period. This removes the dependency of achievement gains on achievement at the start of the period (Cohen, Cohen, West, & Aiken, 2003).[5] This model is considered the base model in that the control covariates are limited to what is typically recommended as the minimal controls for VAA. Comparing the null and base model results is instructive because recent research suggests that prior achievement is the most critical control for reducing selection biases in VAA (Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2008). However, it is unclear whether controlling for prior achievement is also critical for addressing summer effects on VAA estimates.

The demographics model adds eleven student background and demographic variables, six classroom demographic composition variables, and six school demographic composition variables to the base model (see the Appendix Table for list of demographic variables used in this study). The demographic composition variables are included because student composition may be associated with summer learning above and beyond the demographic backgrounds of individual students. For

[4] This study uses the least square (LS) residual rather than the empirical Bayes (EB) residuals. While both have been used for VAA, some recent research comparing the efficacy of LS and EB estimators for classifying teachers found few advantages of the EB approach and that LS generally performed as good or better in simulation studies (Guarino et al., 2015; Schochet & Chiang, 2013).

[5] This model is equivalent to a model with achievement measured at the end of the gain period as the outcome and achievement measured at the beginning of the gain period as a covariate in the model (i.e., a covariance adjustment or ANCOVA set-up). The model fit, variance components, and coefficients will be identical for the two approaches, with the exception of the coefficient for the achievement at the start of the period. The advantage of the approach used in the current application is that it allows a sequential model-building process, whereas the gain score outcome is used throughout and a comparison of the null and base models addresses whether controlling for the association between achievement gains and prior achievement reduced summer effects on VAA estimates.

example, if a school intakes an inordinately high percentage of students with demographic characteristics correlated with negative summer achievement trajectories, the instructional progression at the school may need to be altered to accommodate a more predominate summer setback, which may result in a smaller average achievement gain during the school year.

The context model includes all demographic model variables plus the nine additional measures of classroom context and ten additional measures of school context. The additional variables measure aspects of the educational context that previous research suggests are associated with student learning and may also be associated with summer effects. For example, the contextual variable "proportion new" measures the proportion of the students who transfer into the school after the start of the school year. Recent research suggests that such transfer students tend to disrupt the learning environment (Palardy, 2015). Moreover, the rate at which students transfer in after the start of the school year is arguably more of a proxy of neighborhood instability than a measure of teacher or school effectiveness, and if so is likely to be associated with summer effects.

Model specification. The multilevel equations for the context model are shown below. Note that the other models are reduced forms of the context model for which sets of covariates are omitted. The level one (student) equation is:

$$\begin{aligned}\text{Achievement Gains}_{ics} = \pi_{0cs} + \pi_{1cs}\text{Prior Achievement} + \pi_{2cs}\text{Time Adjustment} + \\ \pi_{3cs}\text{SummerSchool} + \sum_{p=4}^{14}\pi_{pcs}\text{Demographics}_p + e_{ics}. \qquad e_{ics} \sim N(0,\sigma^2)\end{aligned} \quad (1)$$

As described above, the outcome is YoY or SY gains in reading or math achievement either from spring-to-spring or fall-to-spring, respectively. The subscripts (*i*, *c*, and *s*) denote the nested structure of the data; students (*i*) are nested in classrooms (*c*), which are nested in schools (*s*). The model controls for prior achievement, the time duration between the test administrations, and whether the child attended summer school. For the YoY outcome, spring of kindergarten achievement test scores are the prior achievement control, whereas for the SY outcome, fall of first grade scores are used. In addition, a set of eleven student (mostly demographic) background control variables are included (see the Appendix Table for descriptions of the variables used in this study). To adjust VAA estimates for differences in student inputs, continuous control variables are grand mean centered, while dummy variables are uncentered. All slope coefficients are fixed. π_{0cs} represents the conditional mean of the outcome for each *c* classroom. e_{ics} represents the student residuals, which describes the deviation in each child's achievement gains compared to the mean gain of the classroom of which the student is a member. σ^2 is the estimated variance of the student residuals in the population.

The level two (classroom) equations are:

$$\begin{aligned}&\pi_{0cs} = \beta_{00s} + \sum_{p=1}^{6}\beta_{0ps}\text{Demographics}_p + \sum_{p=7}^{15}\beta_{0ps}\text{Context}_p + r_{0cs} \qquad r_{0cs} \sim N(0,\tau_\beta)\\ &\pi_{1cs} = \beta_{10s}\\ &.\\ &.\\ &.\\ &\pi_{14cs} = \beta_{140s}.\end{aligned} \quad (2)$$

Conditional classroom mean achievement gains (π_{0cs}) in reading or math are the outcomes. β_{00s} represents the conditional mean on the outcome for each *s* school. r_{0cs} represents the classroom residuals, which describe the deviation in the adjusted mean achievement gains for each classroom from the mean classroom achievement gains of the school.[6] These residuals are also the teachers' value-added estimates. τ_β is the variance in the classroom residuals and describes the variance in achievement gains among classrooms within schools.

The level three (school) equations are:

$$\begin{aligned}
&\beta_{00s} = \gamma_{000} + \sum_{q=1}^{6}\gamma_{00q}\text{Demographics}_q + \sum_{q=7}^{16}\gamma_{00q}\text{Context}_q + u_{00s} \qquad u_{00s} \sim N(0, \tau_\gamma) \\
&\beta_{01s} = \gamma_{010} \\
&. \\
&. \\
&. \\
&\beta_{015s} = \gamma_{013} \\
&\beta_{10s} = \gamma_{100} \\
&. \\
&. \\
&. \\
&\beta_{140s} = \gamma_{1400}.
\end{aligned} \tag{3}$$

The outcome is conditional reading or math achievement gains at each school (β_{00s}). The intercept, γ_{000}, is the adjusted grand mean achievement gains. The school model includes two sets of covariates: six measures of school demographics and ten measures of school context. All of these covariates are grand mean centered. The school residuals (u_{00s}) represent the deviation in the adjusted mean achievement gains of each school from the grand mean of achievement gains. This residual is also the school value-added estimate. τ_γ represents the estimated variance in the school residuals.

It is worth noting that in comparing YoY and SY VAA estimates, the research design used in the study factors out many conditions that potentially confound such comparisons, including that the same sample of children, teachers, and schools are used for the YoY and the SY estimates, and estimates are based on the same achievement test batteries. The only difference is whether summer is included. This strengthens the internal validity of the design for making inferences about summer effects.

---

[6] Beyond the standard assumptions for general linear models of independent and identically distributed residuals, the viability of inferences from VAA estimates requires several additional assumptions. Some of the additional assumptions stem from the implied causal effects of VAA (i.e., VAA purport to estimate the unique contributions of individual teachers and schools to students' achievement). Reardon and Raudenbush (2009) outline six additional assumptions: manipulability, no interference between units, interval scale metric, homogeneity of effects, strongly ignorable assignment, and functional form. They conclude that at least three of these are unlikely to be met under typical VAA conditions for school effectiveness and that violations degrade the quality of estimates. However, the quality of the inferences for the present study are less dependent on these additional assumptions because the focus is on the impact of including summer, holding all other factors constant.

Results

Research Question 1

To promote policy relevance, the first question is addressed using estimates from the null and base models because they are specified to be consistent with recently-enacted the federal accountability guidelines. Specifically, to receive a waivers from NCLB accountability regulations, states are forbidden from adjusting for demographics such as race/ethnicity, free or reduced price lunch (FRL), or school composition in their accountability models (US DOE, 2010). We quantify the effect of including summer on VAA estimates using two methods: (a) linear associations, including the correlation and squared correlations (R^2) between YoY and SY VAA estimates; and (b) the quintile ranking differences of YoY and SY VAA estimates. Quintile ranks are of policy relevance because VAA are often used to identify and target low-performing teachers and schools for professional development or other remediation and to recognize high-performing teachers and schools for exemplary status.

YoY-SY correlation and R^2. The null model YoY-SY VAA correlations range from 0.61 for schools on math gains to 0.77 for teachers on reading gains (see Table 1). While these correlations are moderate to strong in an absolute sense, they are rather weak for variables purported to measure the same outcome (i.e., teacher or school performance based on gains in student achievement test scores in reading or math). The R^2 values show this more vividly. The R^2 values indicate that the null model SY VAA estimates account for only 38-60% of the variance in YoY estimates, depending on whether the outcome is reading or math achievement gains. The rest of the variance in YoY VAM estimates, 40-62%, originates from the summer period. Note that the YoY-SY associations tend to be weaker for math than for reading. That pattern was expected because math learning is more school-based than is reading. That is, children across demographic groups tend to have little exposure to math over summer, but children from higher-SES families tend to engage in considerable verbal and some written communications with their more educated parents over summer, which can maintain or even build reading and literacy skills over summer (Burkam, Ready, Lee, & Logerfo, 2004; Cooper et al., 1996).

Compared with the null model, the base model correlations are all higher, now ranging from 0.827 to 0.909, and the R^2 values are substantially higher in some cases, now ranging from 0.68 to 0.83. This indicates that controlling for the dependency between the achievement gain outcome and prior achievement reduces the effect of including summer on VAA estimates. The HLM results (not shown due to space limitations) provide an explanation: controlling for prior achievement accounts for considerable variation in mean classroom and mean school summer achievement gains, but a much smaller proportion of mean classroom and mean school SY gains. Note that the summer period is the difference between YoY and SY VAA estimates. Hence, in controlling for prior achievement and reducing the summer effect, the associations between YoY and SY VAA estimates are strengthened.

Table 1

YoY-SY Correlation and Quintile Rank Difference by Model

Comparison	Null Model		Base Model		Demographics		Context Model	
	Reading	Math	Reading	Math	Reading	Math	Reading	Math
				Teachers				
Linear Associations								
Correlation	0.77**	0.64**	0.90**	0.83**	0.91**	0.84**	0.93**	0.86**
R-square	0.60	0.41	0.81	0.68	0.82	0.70	0.86	0.75
Percent Quintile Rank Differences								
Zero	49.1	38.9	59.4	48.7	59.5	48.7	63.9	52.8
One	36.6	38.7	34.6	39.0	34.9	41.1	32.8	38.4
Two	11.6	15.8	5.7	9.8	5.2	8.3	2.9	7.5
Three	2.0	5.0	0.2	2.0	0.3	1.7	0.2	1.1
Four	0.5	1.6	0.0	0.4	0.0	0.1	0.0	0.1
Mean	0.68	0.92	0.47	0.67	0.46	0.64	0.40	0.58
				Schools				
Linear Associations								
Correlation	0.78**	0.61**	0.91**	0.84**	0.89**	0.82**	0.91**	0.84**
R-square	0.60	0.38	0.83	0.70	0.80	0.68	0.83	0.71
Percent Quintile Rank Differences								
Zero	39.3	42.3	52.4	46.4	56.5	51.2	58.3	51.2
One	47.0	38.1	43.5	44.6	35.7	39.3	36.9	42.2
Two	11.3	14.2	3.6	8.4	7.2	8.4	4.2	4.8
Three	1.8	3.6	0.6	0.6	0.0	1.2	0.0	1.8
Four	0.6	1.8	0.0	0.0	0.6	0.0	0.6	0.0
Mean	0.77	0.85	0.52	0.63	0.52	0.60	0.48	0.57

Percent quintile rank differences describe the percent of teachers or schools whose YoY and SY VAA ranks differ by zero, one, two, three, and four quintiles, where zero indicates no difference. For example, the null model results for schools on the reading gains outcome show that 39.3% of the schools have the same YoY and SY rank, while 0.6% differ by four quintiles. The mean is the average YoY-SY quintiles difference. For example, in a sample of two schools, if one school has no quintile difference and the other school has a two quintile difference, the mean is 1.00.

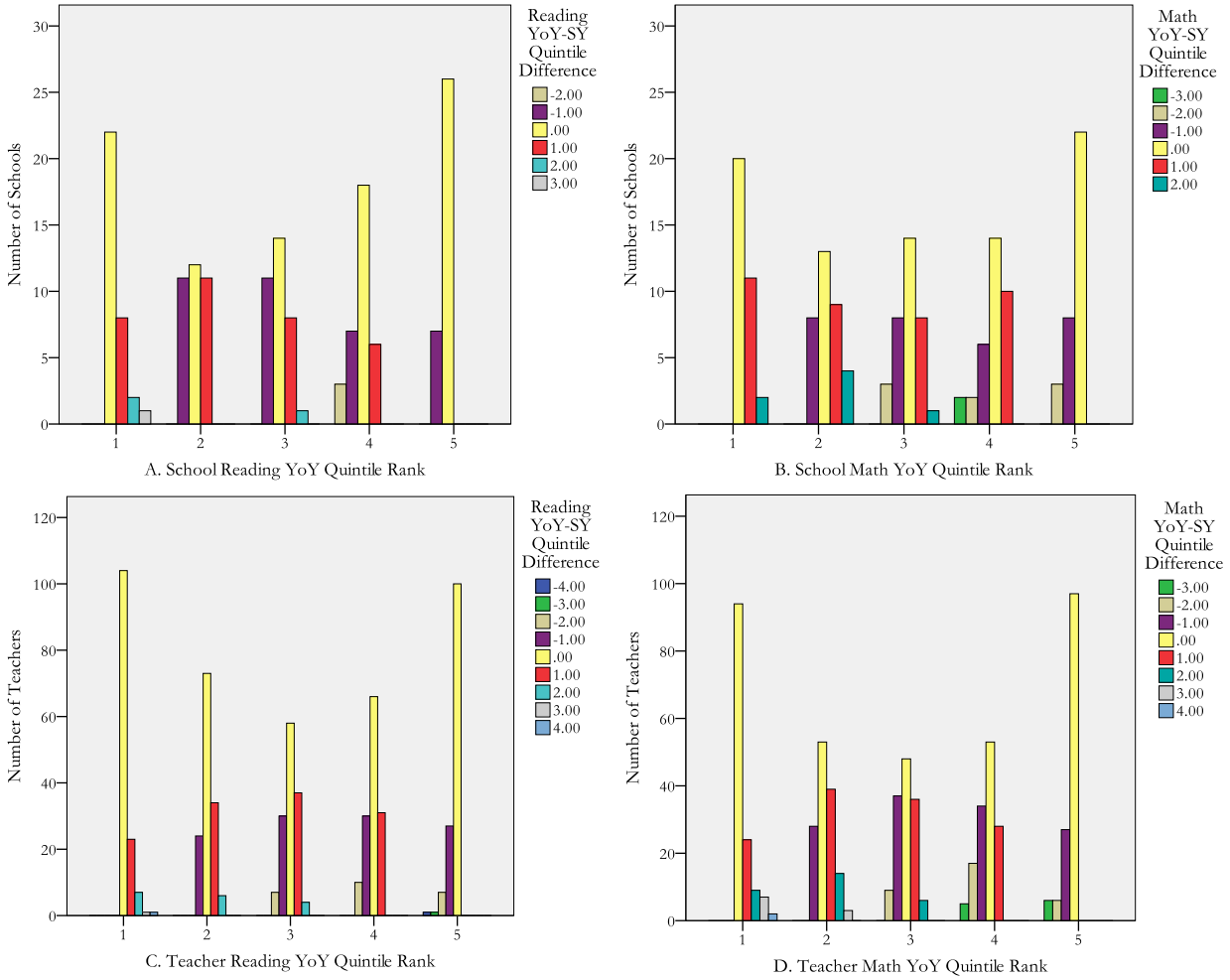
YoY-SY quintile rank differences. The results of the null model quintile comparisons indicate that a large percentage of the teachers and schools are in different effectiveness quintiles for YoY and SY VAA estimates (see Table 1). Between 50.9% and 61.1% of the teachers and schools were in different YoY and SY performance quintiles, depending on whether the outcome was reading or math achievement gains. Moreover, 13.7% to 22.4% differed by two or more quintiles, with several differing by 3 and even 4 quintiles.[7] These differences in YoY-SY quintile ranks are solely due to whether the summer period was included in the achievement gains estimates.

Similar to the results for linear associations, controlling for prior achievement in the base model reduced the quintile rank differences considerably. The magnitude of the reduction can be gauged by comparing the mean quintile rank difference for the null and base models. The mean

[7] Note that a two-quintile difference equates to an average teacher (middle quintile) being classified as very low-performing or very high-performing (quintiles 1 or 5) and a four-quintile difference equates to a very low-performing teacher (quintile 5) being classified as very high-performing (quintile 1) or vice-versa.

quintile differences for teachers on the reading and math gains outcomes were reduced by 31% (from 0.68 to 0.47) and 27% (from 0.92 to 0.67), respectively. The reductions for schools were highly similar. However, even after controlling for prior achievement, between 40.6% and 51.3% of the teachers and schools were still in different quintiles for YoY and SY VAA estimates, and between 4.2% and 12.2% differed by two or more quintiles, depending on the outcome.

Figures 1a-d show that there is approximately the same number of positive and negative quintile misclassifications. That is, teachers and schools are approximately as likely to be underestimated based on YoY VAA quintile rankings as they are to be overestimated. Note, however, that teachers and schools in the lowest YoY quintile can only be classified in equal or higher on SY quintile rank because there is no possibility of being in a lower quintile. Similarly, teachers and schools in the highest YoY quintile are systematically lower on SY rank. It follows that the misclassification rate is highest in the middle YoY quintiles because the teachers and schools can be classified higher or lower on SY. That is, teachers and schools in YoY quintile 3 are more likely to be misclassified than are teachers and schools in YoY quintiles 1 or 5.

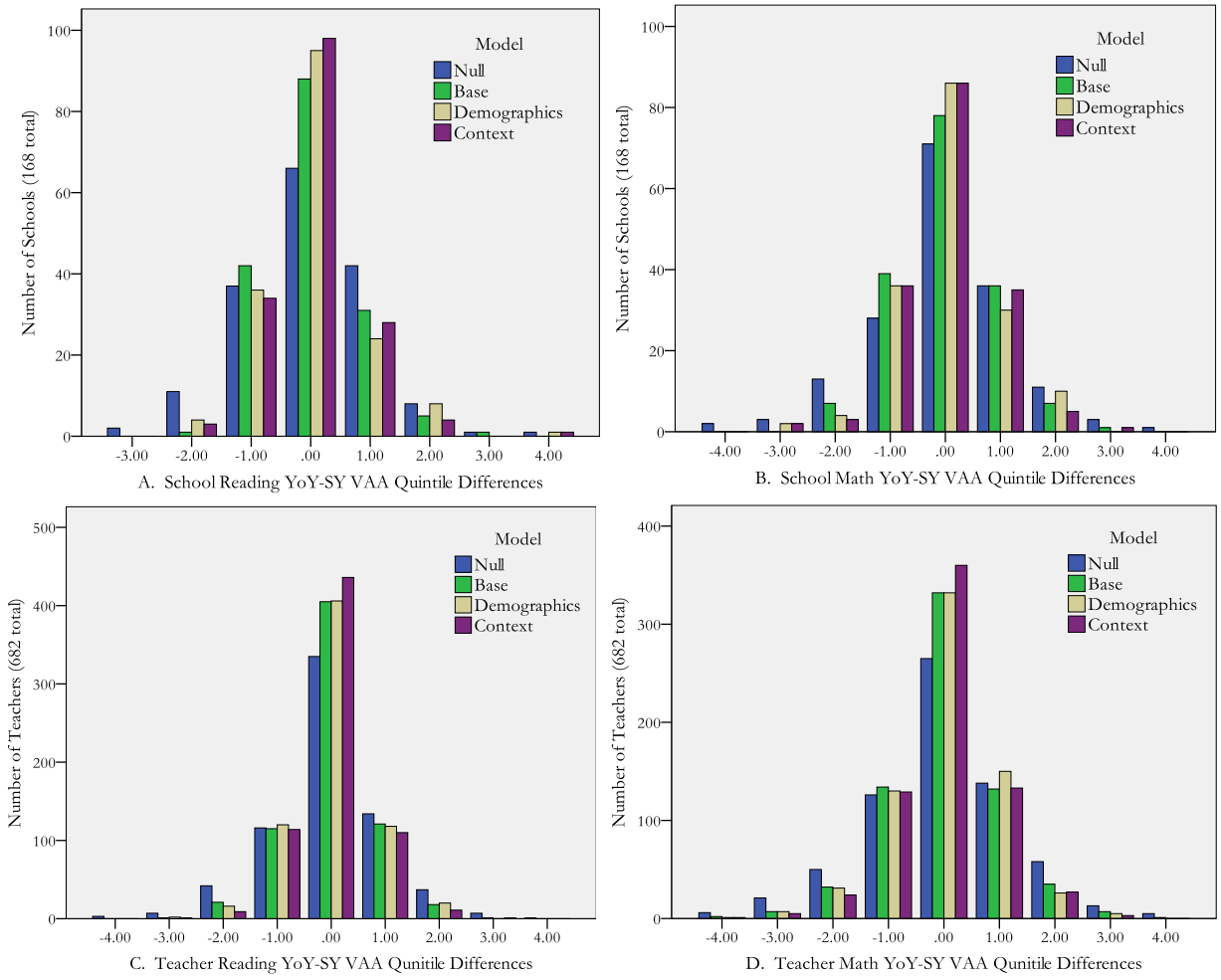


Figures 1a-d: Base Model Quintile Misclassification by YoY Quintiles.

Research Question 2

The purpose of this question is to determine whether summer effects on VAA estimates can be ameliorated using control covariates that are predictive of summer learning, or if biannual assessments (fall and spring) are necessary. To address this, two additional sequential models were fit, including the demographics model and context model (described above). Relevant to policy, the measures included in these models are typically available to districts and thus can be implemented in VAA.

The results (see Table 1 and Figure 2a-d) show that compared with the base model, the demographics model provides only minor improvements in terms of the strength of the linear association between YoY and SY VAA estimates and differences in quintile rank. Similarly, compared with the demographics model, the context model reduced the linear association between YoY and SY only slightly and the differences in quintile rankings are minor. Therefore, including an extensive number of demographic and contextual variables does not substantively reduce the summer effects on VAA estimates. Moreover, after controlling for these extensive sets of variables, substantial YoY-SY quintile rank differences remain. These results suggest that twice-annual assessments (fall and spring) may be necessary to remove the summer effects from VAA estimates.



Figures 2a-d: School and Teacher YoY-SY Quintile Rank Differences by Model.

Research Question 3

To address this research question, the summer part of the base model YoY VAA estimates was isolated from the school-year part. Again, the base model was used because it conforms to the new federal accountability waiver provision that forbids adjustments for demographics (US DOE, 2010). The summer part was isolated by regressing the base model YoY VAA estimates on the SY estimates and saving the model residuals. That was done for teachers and schools separately. These summer VAA effects were then regressed on two measures of student composition that may bias VAA estimates against teachers and schools serving disadvantaged populations: 1) the proportion of students in the classroom or school who receive FRL, and 2) the proportion of students who are black or Hispanic.

The base model results (see Table 2) show no biases among teachers in the same schools. This was expected because first grade teachers in the same school tend to serve highly similar students in terms of students' economic and ethnic backgrounds. However, a significant negative association was found between mean summer gains and the proportion of students at the school who quality for FRL (reading gains = -0.29, $p < 0.01$; math gains = -0.18, $p < 0.05$). Whether proportion FRL and proportion minority are associated with summer biases was also tested for the demographics model. Recall that the demographics model controls for those and other demographics factors, so no biases were expected. The results (see Table 2) confirm that.

Table 2
The association between the summer component of VAA estimates and proportion underserved students in the classroom or school.

	Base Model		Demographics Model	
	Reading	Math	Reading	Math
Classroom				
Proportion FRL	-0.06	-0.06	0.00	0.01
Proportion Minority	-0.03	-0.01	-0.01	-0.01
School				
Proportion FRL	-0.29**	-0.18*	0.01	0.01
Proportion Minority	-0.03	-0.02	-0.03	-0.01

Coefficients are in units of standardized effect size. ** $p<0.01$, * $p<0.05$;

Discussion

The results for research question 1 show that a substantial portion of the variance in YoY VAA estimates originates from summer and that summer variance alters the quintile rankings of a high percentage of the teachers and schools. These findings are consistent with the two previous studies on effects of test timing on VAA, one with implications to teacher VAA and the other with implications to school VAA. Papay (2011) concluded that test timing, such as whether the test was administered in fall and spring and whether the test was administered very close to the end of the school year (i.e., May) or earlier (i.e., March), was the largest source of measurement error on VAA of teacher effectiveness. Downey et al. (2008) estimated 20-35% of the schools that were classified in the bottom or top quintile when summer was included in the VAA model were ranked in another quintile when summer was excluded.

The findings for research question 3 show that including summer in VAA estimates results in systematic biases against schools serving higher concentrations of students who qualify for FRL.

This result is consistent with the literature on seasonal effects, which indicates that students from low SES families tend to have greater declines in reading achievement over summer, but learn at similar rates as other children during the school year (Alexander et al., 2001; Cooper et al., 1996; Heyns, 1978).

Implications to Practices for Reducing Summer Biases

Twice-annual assessments. The findings for research question 2 suggest that addressing summer effects on teacher and school VAA estimates will require twice-annual assessments (fall and spring). That is because even after employing extensive statistical controls for student background and demographics, as well as controls for classroom and school context, substantial differences in YoY and SY VAA estimates remained.

Controlling for prior achievement. A comparison of the results for the null and base models shows that controlling for prior achievement reduces the summer effect considerably. Previous research has shown that controlling for prior achievement reduced selection biases in VAA estimates (Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2008). That conclusion appears to extend to selection biases originating from the summer. Hence, if twice-annual assessments are not conducted, controls for prior achievement seem to be the best method for minimizing summer effects.

Student assignment practices. The results suggest that once enrolled at a school, first graders are not randomly assigned to teachers. That is, students attending the same school are expected to vary in terms of summer learning rates, but if the first graders enrolled at a given school are randomly assignment to their first grade classrooms, then the mean summer learning rates among classrooms in the same school would be expected to exhibit only random variation. Yet, the results show a substantial degree of variation in summer learning rates among classrooms in the same school, suggesting that children are not randomly assigned to classrooms. This finding is not surprising, as previous research has concluded that random assignment of students to classrooms is uncommon (Authors, 2015; Burns & Mason, 1998; Kalogrides & Loeb, 2013; Paufler & Amrein-Beardsley, 2014; Praisner, 2003; Rothstein, 2010). The results for the base model suggest that students' prior achievement plays a role in student assignment because controlling for prior achievement substantially reduced summer effects on teacher VAA. However, the results for the demographics model suggest that demographics play only a very minor role. Hence, other than prior achievement, it is not clear what the precise student placement mechanisms are that contribute to summer effects in VAA teacher estimates.

A recent study by Paufler and Amrein-Beardsley (2014) provides insight into what those student assignment mechanisms might be. The authors surveyed over 300 elementary school principals in Arizona, 98% of whom reported using student and teacher information during the placement process in an effort to match learning and teaching styles, personalities, and special needs with the objective of maximizing student outcomes. The student information that principals reported giving the strongest consideration to was prior academic achievement, prior behavioral issues and/or perceived behavioral needs, language status and/or proficiency, and prior grades. The present study controls for prior achievement and language status, but not behavioral issues and needs or grades, because good measures of those variables were not available for the ECLS data. Research is needed to examine whether student assignment practices that take into account students' behavioral issues/needs or grades contribute to the summer effects on VAA estimates.

Reducing measurement error. The high rates of quintile rank differences between YoY and SY VAA indicate that including summer adds considerable measurement error to VAA

estimates, which undermines their reliability.[8] Previous research on VAA has shown that teacher—and to a lesser extent, school—VAA are unreliable from year to year; however, those studies did not examine the degree to which including summer contributed to the unreliability (Lockwood et al., 2007; Papay, 2011). Similarly, previous research has shown that the reliability of teacher VAA estimates can be improved substantially by pooling data across multiple years (see McCaffrey, Sass, Lockwood, & Mihaly, 2009). However, it is not clear whether pooling data across multiple years will address the summer effects. Pooling data across years improves reliability of VAA estimates by accounting for year-to-year fluctuations due to measurement error and other random factors, as well as year-to-year fluctuations in true performance. Yet, if the summer effect is based on the same mechanisms across years (e.g., student assignment practices), pooling the data across years will not likely reduce it. Research is needed to determine the degree to which summer effects are consistent across years and whether twice-annual testing addresses the more general issue of year-to-year instability among VAA estimates.

Policy Implications

Cost-benefits of twice-annual assessments. The results of this study have important implications for educational policy regarding the inclusion of the summer period in VAA. Perhaps the most critical implication is that fully addressing summer effects will likely require twice-annual achievement testing. However, such a proposal may be met with opposition due to concerns about costs associated with additional testing and the time it would take from learning activities. Yet, the validity of the cost concern is questionable. For example, a recent Brookings Institute study found that achievement test batteries cost an average of $27 per pupil in grades 3-9, which represents a miniscule percentage of total annual per-pupil expenditures (Chingos, 2012).[9] In addition, the study concluded that the already low costs of testing can be reduced by a third or more if states participate in a testing consortium such as Smarter Balanced Assessment, which distributes test development and scoring expenses across an extremely large number of students. With the onset of Common Core, most states have recently joined a testing consortium already. Therefore, concern about additional expenses is not a good reason for rejecting biannual testing.

A more realistic concern than the monetary expenses associate with additional testing is the time it will take from learning activities. Standardized achievement test batteries typically take 3-8 hours to administer. Furthermore, when stakes are attached to the results, time may be spent on test-specific preparation that is of questionable value to academic development. Given the results of this study, federal and state agencies should consider policies that encourage exploration of the cost-benefits of twice-annual testing. Critical to that analysis is a better understanding of how much time an additional annual test battery is expected to take from learning activities and whether that can be reduced.

High-stakes personnel decisions. The results show that when summer is included, VAA model estimates can be very different compared to when summer is not included. This raises concerns about the use of YoY VAA estimates for high-stakes personnel decisions. While an argument can be made that YoY VAA estimates still contain useful information about the

---

[8] A note of caution is in order here: while a strong argument can be made that SY gains are a more valid outcome measure for VAA than YoY gains, SY gains may also include some measurement error originating from summer. For example, if students whose achievement declines most over summer tend to rebound most during fall, the rebound effect may have little to do with teacher or school performance.

[9] This estimate is similar to an inflation-adjusted estimate from 1993 by the non-partisan U.S. General Accounting Office (GAO) of $24 to $53 per pupil. The high boundary of the GAO estimate assumes the tests are administered by hired external personnel, which is uncommon for standardized achievement tests.

performance of individual teachers or schools (e.g., see Glazerman et al., 2010), their marginal reliability suggests they should not be the sole basis for gauging performance for high stakes decisions.

Biases against schools serving disadvantaged children. Another finding of this study with policy implications is that VAA estimates are biased against schools serving higher percentages of children who quality for FRL. This summer effect can easily be addressed by controlling for differences among schools in the proportion of students who receive FRL. However, recent federal policy on accountability waivers forbids the use of such controls (US DOE, 2010). The results of this study challenge the fairness of that policy, suggesting that it will result in systematic bias against high-poverty schools, which can create false perceptions that such schools and the teachers and administrators working there are ineffective, when their performance is average or even above average. Biased VAA estimates and the perceptions they create can have negative consequences on staff morale and efforts to recruit and retain effective teachers and administrators.

Limitations and Future Research

A limitation of this study is that the results are based on data from first grade. The reason for that limitation is that ECLS-K only has spring-fall test scores for one year—between kindergarten and first grade. It is unclear whether the results of this study generalize to higher grade levels. However, due to age proximity and similarities in instructional methods and classroom structure in early elementary school, the results may generalize to second and third grades. Research is needed to examine summer effects on VAA at higher grade levels.

Another limitation is the size of the classroom sample.[10] The average classroom sampled had 3.3 students. Having data on all students in each classroom would improve the reliability of the individual teacher VAA estimates. However, it is not clear how this affects VAA quintile misclassification rates. To examine that, a sensitivity analysis was conducted. The analysis used a subsample of teachers who had the largest number of children in their sample. The cut-points for being included in the sensitivity analysis were teachers with 8 or more students sampled (n = 31). The results for this analysis (see Table 3) were compared with the results for the full sample (Table 1), shows consistency in misclassification rates. There is no evidence that sample size impacts misclassification rate in a systematic manner. The results of this sensitivity analysis are not surprising because this study does not examine the reliability of VAA estimates per se, but rather the impact of summer on VAA misclassification. These are two different issues, with the former highly impacted by sample size and the latter apparently much less so.

It is also worth noting while a substantial number of control variables were used to test whether the summer effects were due to student demographics or classroom and school context, other types of variables may have contributed to summer effects. The control variables used in this study were selected for two reasons: 1) they are typically available to school personnel and therefore can readily be implemented in accountability models; and 2) previous research suggests they are associated with summer effects. Research is needed to examine whether other types of control variables, such as summer activities and neighborhood effects, can reduce summer effects on VAA estimates.

[10] Note that all first grade classrooms in each school were used and therefore the number of classrooms per school cannot be increased unless multiple grade-levels are pooled together.

Table 3
Sensitivity analysis of classroom and school sample size

Comparison	Null Model		Base Model		Demographics		Context Model	
	Reading	Math	Reading	Math	Reading	Math	Reading	Math
Teachers with 8 or more students in sample (n=31)								
Percent Quintile Rank Differences								
Zero	45.2	41.8	54.8	48.4	54.8	51.6	54.8	51.6
One	51.6	35.5	45.2	41.9	45.2	41.9	45.2	48.4
Two	3.2	19.4	0.0	9.7	0.0	6.4	0.0	3.2
Three	0.0	3.2	0.0	0.0	0.0	0.0	0.0	0.0
Four	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sum	58.1	83.9	45.2	64.5	45.2	61.3	45.2	54.8

Summary and Conclusions

The findings of this study show that between 40% and 62% of the variance in YoY VAA estimates, depending on the outcome, originates from the summer period. This summer measurement error alters teacher and school quintile rankings considerably. For example, 51% to 61% of the teachers and 58% to 61% of the schools, depending on the outcome, change performance quintile rank when the summer period is omitted, and many teachers and schools change 2 to 3 quintiles and a few changing 4 quintiles. Furthermore, this summer effect invariably underestimated the performance of teachers and schools in the lowest quintile of summer change and overestimated the performance of teachers and schools in the highest quintile of summer change. While controlling for prior achievement reduces the YoY-SY VAA differences, extensive statistical controls for student background, demographics, and classroom and school context did not substantially alter the summer effect. Finally, including the summer period in VAA estimates created biases against schools serving high concentrations of children who qualify for FRL, and while statistical controls can neutralize those biases, current federal policy forbids their use for accountability assessments. Together, these findings indicate that including summer in VAA substantially undermines the reliability of VAA estimates and that addressing the problem will likely require biannual (fall and spring) achievement testing.

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. http://dx.doi.org/10.3102/01623737023002171

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*, 65-75. http://dx.doi.org/10.3102/0013189X08316420

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R.,…Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper #278. Washington, DC: Economic Policy Institute.

Betts, J. R., Zau, A. C., & Rice, L. A. (2003). *Determinants of student achievement: New evidence from San Diego.* San Francisco, CA: Public Policy Institute of California.

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

Braun, H., Chudowsky, N., & Koenig, J. (Eds.) (2010). *Getting value out of value-added: Report of a*

workshop. Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability; National Research Council. http://www.nap.edu/catalog/12820.html

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. http://dx.doi.org/10.1177/003804070407700101

Burns, R., & Mason, D. (1998). Class formation and composition in elementary schools. *American Educational Research Journal, 35,* 739-772. http://dx.doi.org/10.3102/00028312035004739

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*, 2593-2632. http://dx.doi.org/10.1257/aer.104.9.2593

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed). Hillsdale, NJ: Erlbaum.

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. http://dx.doi.org/10.3102/00346543066003227

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*, 242–270. http://dx.doi.org/10.1177/003804070808100302

Glazerman, S., Loeb, S., Goldhaber, D., Raudenbush, D., Staiger, D., & Whitehurst, G. J. (2010). Evaluating teachers: The important role of value-added. Washington, DC: The Brookings Brown Center.

Guarino, C., Maxfield, M., Reckase, M., Thompson, P., & Wooldridge, J. (2015). An evaluation of empirical Bayes estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics. 40,*190-222. http://dx.doi.org/10.3102/1076998615574771

Guarino, C., Reckase, M., & Wooldridge, J. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy, 10*, 117-156. http://dx.doi.org/10.1162/EDFP_a_00153

Haertel, E. H. (2013). Reliability and validity of inferences about teachers based on student test scores (14th William H. Angoff Memorial Lecture). Princeton, NJ: Educational Testing Service. Available at http://www.ets.org/Media/Research/pdf/PICANG14.pdf

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Henry, G. T., Rose, R. A., & Lauen, D. L. (2014). *Are value-added models good enough for teacher evaluations? Assessing commonly used models with simulated and actual data.* Paper presented at the Economics of Education Association Annual Conference. Downloaded from http://2014.economicsofeducation.com/user/pdfsesiones/170.pdf?PHPSESSID=mpvim1aqj4urod74ugvvovb202

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

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher, 42*, 304-316. http://dx.doi.org/10.3102/0013189X13495087

Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.* NBER Working Paper 14607.

Koedel, C. & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy, 5*, 54-81. http://dx.doi.org/10.1162/edfp.2009.5.1.5104

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*, 47–67. http://dx.doi.org/10.1111/j.1745-3984.2007.00026.x

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND. http://dx.doi.org/10.1037/e658712010-001

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 37-66. http://dx.doi.org/10.3102/10769986029001067

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. http://dx.doi.org/10.1162/edfp.2009.4.4.572

Murnane, R. J., Willett, J. B., Bub, K. L., & McCartney, K. (2006). Understanding trends in the Black-White achievement gaps during the first years of school. *Brookings-Wharton Papers on Urban Affairs, 7*, 97–135. http://dx.doi.org/10.1353/urb.2006.0024

National Center for Educational Statistics (2002). *User's guide to the longitudinal kindergarten-first grade public-use data file*. Washington, DC: U.S. Department of Education.

Palardy, G. J. (2010). The multilevel crossed random effects growth model with applications for estimating teacher and school effects: Issues and extensions. *Educational and Psychological Measurement, 70,* 401-419. http://dx.doi.org/10.1177/0013164409355693

Palardy, G. J. (2015). The effects of classroom context, teacher qualifications and effectiveness on achievement gaps in first grade. *Teachers College Record, 117*, 1-48.

Palardy, G. J., & Rumberger, R. W. (2008).Teacher effectiveness in the first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis, 30,* 111–140. http://dx.doi.org/10.3102/0162373708317680

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Education Research Journal, 48*, 163-193. http://dx.doi.org/10.3102/0002831210362589

Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal, 51*, 328-362. http://dx.doi.org/10.3102/0002831213508299

Praisner, C. (2003). Attitudes of elementary school principals toward the inclusion of students with disabilities. *Exceptional Children, 69*, 135-145. http://dx.doi.org/10.1177/001440290306900201

Raudenbush, S. W. (2013). What do we know about using value-added to compare teachers who work in different schools? Carnegie Knowledge Network. Retrieved from: http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/08/CKN_Raudenbush-Comparing-Teachers_FINAL_08-19-13.pdf

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating

school effects. *Education Finance and Policy, 4*, 492-519. http://dx.doi.org/10.1162/edfp.2009.4.4.492

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics, 125*, 175-214. http://dx.doi.org/10.1162/qjec.2010.125.1.175

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics, 38*, 142-171. http://dx.doi.org/10.3102/1076998611432174

Stipek, D. (2004). Teaching practices in kindergarten and first grade: Different strokes for different folks. *Early Childhood Research Quarterly*, *19*, 548–568. http://dx.doi.org/10.1016/j.ecresq.2004.10.010

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J.,…Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics 29*, 11-36. http://dx.doi.org/10.3102/10769986029001011

U.S. Department of Education (2010). *Interim report on the evaluation of the growth model pilot project*. Retrieved from: http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp.pdf

Appendix A
Variable Descriptions

Variable Name	Mean (SD)	Description (ECLS variable label)
Student Variables (Level-1, n = 2,251)		
Achievement Change Outcomes		
Year-over-year Reading	0.83 (0.37)	Spring 1st minus spring K (c4r4thtr-c2r4thtr)*
Year-over-year Math	0.75 (0.32)	Spring 1st minus spring K (c4m4thtr-c2m4thtr)*
School-year Reading	0.84 (0.38)	Spring 1st minus fall 1st (c4r4thtr-c3r4thtr)*
School-year Math	0.69 (0.35)	Spring 1st minus fall 1st (c4m4thtr-c3m4thtr)*
Summer Reading	-0.01 (0.26)	Fall 1st minus Spring K (c4r4thtr-c3r4thtr)*
Summer Math	0.06 (0.29)	Fall 1st minus Spring K (c4m4thtr-c3m4thtr)*
Time Measurement Adjustments		
Year-over-year	12.00 (0.00)	Months btw end of kindergarten and end of 1st*
School-year	9.44 (0.32)	Months btw start of 1st grade and end of 1st*
Summer	2.56 (0.32)	Months btw end of K and start of 1st*
Prior Achievement Controls		
Spring K Math	-0.57 (0.44)	Spring K reading score (c2m2tht_r)
Spring K Reading	-0.60 (0.49)	Spring K math score (c2r2tht_r)
Fall 1st Grade Math	-0.51 (0.46)	Fall 1st grade reading score (c3m3tht_r)
Fall 1st Grade Reading	-0.61 (0.49)	Fall 1st grade math score (c3r3tht_r)
Spring 1st Grade Math	0.18 (0.40)	Spring 1st grade reading score (c4m4tht_r)
Spring 1st Grade Reading	0.23 (0.43)	Spring 1st grade math score (c4r4tht_r)
Demographic and Background Controls		
Free or Reduced Lunch	0.36	Parent states child receives FRL (p4lunchs)
Female	0.48	(gender = 1) recoded to 0 = male, 1 = female
Asian	0.03	(race = 5)
Black	0.14	(race = 2)
Hispanic	0.14	(race = 3 or 4)
Other	0.06	(race = 6, 7, or 8)
White (reference group)	0.62	(race = 1)
Age (months)	79.93 (4.34)	Age in months at fall of 1st grade (R3AGE)
LEP	0.08	Non-English home language (WKLANGST=1)
Disability	0.14	Parent states child has a disability (P1DISABL)
Days Absent	8.11 (7.54)	Total days absent during 1st grade (U4ABSN)
Summer School	0.10	Attended summer school (P3SUMSCH)
Classroom Variables (Level-2, n = 682, Mean Classroom Sample Size = 3.4)		
Classroom Demographics (classroom means)		
Proportion Free Lunch	0.40 (0.39)	Proportion free or reduced lunch (mean p4lunchs)
Proportion Minority	0.32 (0.39)	Proportion Latino or black (a4pmin/100)
Proportion Female	0.48 (0.33)	Proportion female (mean gender = 1)
Proportion Disability	0.14 (0.23)	Percent of students with disability (a4disab/a4totag)
Proportion LEP	0.10 (0.25)	Percentage LEP students in class (a4numle/a4totag)
Mean Age	80.08 (3.17)	Mean age in months fall of 1st grade (mean R3AGE)

Appendix Table (Cont'd.)
Variable Descriptions

Variable Name	Mean (SD)	Description (ECLS variable label)
Classroom Context (classroom means)		
Proportion New	0.17 (0.25)	Percent new students (a4new/a4totag)
Mean Days Absent	8.42 (6.18)	Days absent during 1st grade (U4ABSN mean)
Mean Math	-0.54 (0.36)	Classroom math achievement (c3m3thtr mean)
Mean Reading	-0.63 (0.39)	Classroom reading achievement (c3r3thtr mean)
Math Heterogeneity	0.35 (0.18)	Classroom var in math (var c3m3thtr)
Reading Heterogeneity	0.38 (0.20)	Classroom var in reading (var c3r3thtr)
Large	0.11	More than 25 students in class (a4totag > 25)
Small	0.09	Fewer than 17 students in class (a4totag < 17)
Proportion Summer School	0.11 (0.23)	Proportion summer school (P3SUMSCH mean)
School Variables (Level-3, n = 168, Mean Number of Classrooms per School = 4.1)		
School Demographics (school means)		
Proportion FRL	0.43 (0.31)	Proportion FLR (mean p4lunchs)
Proportion Minority	0.35 (0.34)	Proportion Minority (mean (a4pmin/100))
Proportion Female	0.47 (0.14)	Proportion female (mean gender = 1)
Proportion Disability	0.13 (0.10)	Proportion disability (mean a4disab/a4totag)
Mean Proportion LEP	0.12 (0.20)	Proportion LEP (mean a4numle/a4totag)
Mean Age	80.05 (1.98)	Age in months in fall of 1st grade (mean r3age)
School Context (school means)		
Mean Proportion New	0.19 (0.18)	Proportion new students (mean a4new/a4totag)
Mean Days Absent	8.33 (3.36)	Days absent during 1st grade (mean U4ABSN)
Mean Math	-0.53 (0.23)	Mean math achievement (mean c3mrscal)
Mean Reading	-0.63 (0.26)	Mean reading achievement (mean c3rrscal)
Math Heterogeneity	0.41 (0.10)	Variance math achievement (var c3m3thtr)
Reading Heterogeneity	0.43 (0.12)	Variance reading achievement (var c3r3thtr)
Proportion Large	0.10 (0.26)	Proportion large classrooms (mean a4totag > 25)
Proportion Small	0.10 (0.27)	Proportion small classrooms (mean a4totag < 17)
School Safety	0.00 (1.00)	Principal component
Proportion Summer School	0.10 (0.13)	Summer school attendance (P3SUMSCH)

Notes: All student variables are weighted using the student sampling weight provided by NCES (C34CW0). *The spring kindergarten (K) assessment was administered 1-4 months before the start of summer, and the fall 1st grade assessments were administered 1-3 months after the start of 1st grade. To compute an accurate estimate of summer change, the spring K assessments were extrapolated forward to the start of summer assuming a linear rate of change during K, and the fall 1st grade assessments were extrapolated back to the start of 1st grade assuming a linear rate of change during 1st grade. The adjusted test scores and time measures are provided on this table.

About the Authors

Gregory J. Palardy
University of California, Riverside
gregory.palardy@ucr.edu
Dr. Palardy is on the faculty of the Graduate School of Education at the University of California, Riverside. His research focuses on teacher and school effectiveness with an emphasis on understanding how educational practices, policies, and contexts contribute to student outcomes and to educational opportunity. Recent studies have examined the effects of socioeconomic and ethnic segregation in schools on student attainment, learning, and school behaviors; the effects of inequitable access to qualified and effective teachers on achievement gaps; and the effects of social mechanisms in schools, such as peer influences, social capital, and organizational habitus, on student outcomes.

Luyao Peng
University of California, Riverside
luyao.peng@email.ucr.edu
Luyao Peng is a doctoral student in Educational Psychology and a Master's student in Applied Statistics. Her research interests are in educational measurement and testing and applications of multilevel and item response theory models. She is currently studying the efficacy of the deterministic gated IRT model for detecting cheating on test items.

education policy analysis archives

Volume 23 Number 92 September 28th, 2015 ISSN 1068-2341