Abstract Title Page
Not included in page count.

Title: The Misattribution of Summers in Teacher Value-Added

Authors and Affiliations: Allison Atteberry, University of Virginia

Abstract Body

Limit 4 pages single-spaced.

Background / Context. *Description of prior research and its intellectual context.*

This paper investigates the extent to which spring-to-spring testing timelines bias teacher value-added as a result of conflating summer and school-year learning. Using a unique dataset that contains both fall and spring standardized test scores, I examine the patterns in school-year versus summer learning. I estimate value-added based on traditional spring-to-spring data, as well as competing models that predict fall-to-spring test score gains. I examine whether teachers are ranked differently using these two testing timelines and whether certain kinds of teachers are especially affected by the test timing. The paper discusses whether this problem is of sufficient magnitude to caution against using value-added measures based solely on spring-to-spring test score data. Since states currently do not require both fall and spring testing, the implications for the federal accountability planning are significant.

If all students learn at the same rate during the summer, then spring-to-spring testing would not systematically bias estimated teacher effects. However, research on summer impacts to student learning have shown that middle-class children exhibit gains in reading achievement over summer, while disadvantaged children showed losses (Alexander, Entwisle, & Olson, 2004). Little is known about precisely what causes students of different socioeconomic and demographic backgrounds to experiences summers so differently, though research has suggested that income differences could be related to differences in opportunities to practice and learn over summer, with more books and reading opportunities available for middle-class children (Alexander, Entwisle, & Olson, 2007; Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; Downey, von Hippel, & Broh, 2004; Heyns, 1978). Because the sources of differences is summer learning are unobserved, value-added models that conflate the summer period with the school year will inappropriately blame teachers with disadvantaged students for this summer decay while the teacher of the advantaged student will be credited for gains they did not foment (see discussion in Scherrer, 2011).

Purpose / Objective / Research Question / Focus of Study. *Description of the focus of the research.*

In recent years, researchers have developed statistical methods for using a new source of evidence—standardized test score data—to estimate teacher effectiveness, despite significant challenges to valid causal inference (for an overview of challenges, see McCaffrey, 2003; Reardon & Raudenbush, 2009). So called "value-added" models typically use district administrative data on students, teachers, and schools in statistical models that seek to isolate the teacher's impact on test score performance from the many other factors that influence student test scores but are outside of the teacher's control.

In practice, however, the estimation of value-added is far from straightforward. The statistical models produce unbiased estimates of teacher contributions to student learning only when they account for all variables that affect both student achievement and differ across teachers. Any such factors that are omitted from the model will be conflated with estimated teacher effects. To address this problem, most value-added models control for commonly observed student characteristics such as prior test scores, gender, race/ ethnicity, language status, special education status, and proxies for socio-economic status such as eligibility for free/reduced price lunch program or parental education. Of course, there remain numerous factors other than the teacher's influence that could reasonably account for any residual changes in student test scores from one year to the next.

Much has been written about statistical challenges to isolating the teacher's causal impact on student test scores, however one obvious problem with typical value-added models has received less attention. Almost all value-added models estimate the effects of teacher using

spring-to-spring outcome data, simply given that statewide annual testing occurs on this timeline. As a result, the teacher is attributed both with learning gains made by students during the school year as well as the prior summer. In essence, teachers are in part evaluated for what happens to students prior to their first day in their class, which runs counter to the fundamental goals of an accountability system.

Setting. *Description of the research location.*

A four-year study beginning in 2004-05 evaluated the efficacy of the Literacy Collaborative (LC), a coach-based intervention program. The project took place in seventeen schools across eight states in the Midwestern, southern and eastern parts of the U.S. Schools were selected for the study from a pool of schools that had previously expressed interest in implementing LC. The schools were selected nationally so that any findings were not specific to a particular local or state context. The LC study concluded in 2007-08 and provided an opportunity to conduct the analysis proposed herein.

Population / Participants / Subjects. *Description of the participants in the study.*

Some 250 K-2 teachers were working in these 17 schools during the study, and most—about 94 percent—participated. Student achievement data was collected in both the spring and fall of all four study-years. The study involved children from six different cohorts who entered at different grades and in different years, and as a result most students have fewer than the complete set of six student test scores (fall/ spring in grades K, 1, 2).

Approximately 1150 students were assessed in each grade in kindergarten through second grade in each year of the study (three grades, four years). This represents a student participation rate of 90 percent or higher at each testing occasion. Only 3 percent of children missed testing on one or more occasions for which they were eligible to be assessed (e.g., due to absences). The LC study sample included four years of data amounting to 42,255 repeated observations of 9,967students attending 250 teachers' classrooms in 17 schools.

Intervention / Program / Practice. *Description of the intervention, program, or practice.*

Teacher value-added models seek to estimate a causal impact of a given teacher on student learning. The "intervention" under investigation here is a set of teachers, and the goal of this paper is to examine the consequences of using annual—rather than fall and spring—test score data to estimate these causal effects.

It is useful to use the Holland/ Rubin causal framework to clarify the goals and challenges of value-added estimation, which is an attempt to capture a causal inference about the effect of experiencing one teacher versus another (Holland, 1986; Rubin, 1974).[2] However the realities of the U.S. school system make it particularly difficult to estimate the potential outcomes of a particular student had he or she experienced one teacher versus another. In the ideal execution of research based on the causal model, the assignment of individual units to treatment status is either completely random or completely observable to the researcher. However, the assignment of students to teachers in the U.S. meets neither of these criteria. Value-added models are imperfect attempts to attend to the challenges of estimating causal impacts of teachers. To date, research has focused on competing specifications of the statistical

[2] According to this framework, a causal effect on some outcome (Y) for some unit (i) of some treatment condition (TT) relative to some other control condition (TC) is defined as the difference between the value of Y that would be observed if unit i were exposed to treatment TT and the value of Y that would be observed if unit i were exposed to treatment TC (Rubin, 1974, p. 689; West, Biesanz, & Pitts, 2000). In the matter of teacher effects, a set of teachers—perhaps within a district—represents the treatments of interest (TT versus TC's) to which students (i) are exposed. The magnitude of the effect of a given teacher is thus estimated by comparing the learning outcomes (Y) of the students who are exposed to these competing treatments.

model and underlying estimation methods, however less attention has been paid to how the limitations of the data used in these models may produce bias. This project thus examines the challenges to estimating the causal impact of teachers given the data constraints that exist as a result of annual testing that typically occurs in spring of each school year.

Research Design. *Description of the research design.*

First, I begin with an exploration of common value-added models used to estimate teacher effects. This includes gain score models, covariate adjustment models, school fixed effects vs. school characteristics, student fixed effects vs. student characteristics, as well as cross-classified growth models. The basic covariate adjustment model is described in more detail in the following section. This section of the paper serves primarily to reproduce findings in other value-added papers that suggest that differences in model specification are relatively small (see, e.g., Papay, 2010).

Next, I present descriptive work on estimated summer learning gains. I estimate the mean summer gain (or loss), as well as the variance. I use the student data available to examine whether any patterns emerge in summer learning. Are certain kinds of students more likely to experience summer learning gains, while others experience loss? How much of the observed variance can be accounted for using the typical student background information available in large-scale administrative datasets (e.g., race/ ethnicity, gender, free-reduced price lunch program eligibility, etc.).

The third key step of the analysis will be to estimate value-added models in which the only difference is the use of current fall versus prior spring test score data to predict current spring test scores. In theory, value-added models that use fall to spring test score data should better isolate the teacher's causal impact, since spring-to-spring test cycles conflate the prior summer with the school year.

Fourth, I quantify the extent of the disagreement between value-added estimates based on these two test-timings. I do so first by calculating the Spearman rank correlation between the results from the two measures. I also rank teachers into quartiles and demonstrate the percentage of teachers who would be ranking in different quartiles based on the two. This provides some evidence about the percentage of teachers who may be inaccurately categorized simply based on on the fact that they are being given credit for (or being held responsible for) the student learning during the summer prior to the school year. I also examine the extent to which different value-added models appear to be more susceptible to this problem. To what extent does the inclusion of traditional student covariates appear to "soak up" the differential summer effects? Though prior research shows that summer learning differs by socioeconomic group, we know less about the true underlying causes of these patterns, and so it is unlikely that summer effects will be controlled for simply by including student race/ ethnicity/ socio-economic status indicators.

Data Collection and Analysis. *Description of the methods for collecting and analyzing data.*

The original researchers used two reading assessments in order to assess broadly students' literacy learning over the primary grades in this study: the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and the Terra Nova Multiple Assessments of Reading.[3] To estimate year-specific teacher effects, I specify a relatively standard teacher-by-year fixed effects

[3] The DIBELS taps a range of early literacy skills, including letter recognition, phonological awareness, decoding, and oral reading fluency (Good & Kaminski, 2002). The Terra Nova is a group-administered, standardized, norm-referenced reading test (McGraw-Hill, 2001). These two tests were scaled together using Rasch modeling (Wright & Masters, 1982). This process generated a vertically-scale test metric to ease interpretation of the longitudinal character of student literacy learning data (Biancarosa, et al., 2010; Kerbow, Biancarosa, Bryk, & Luppescu, 2007). The resulting measures provide time-varying student test score data which serve to generate the value-added measures for teachers over time.

model that includes a lagged prior student test score, a set of time-invariant and time-varying student demographic characteristics, grade fixed effects, and teacher-by-year fixed effects:

$$ElaAch_{igjtk} = aElaAch_{igjt-1k} + X_{igjt}\beta + \theta_g + \delta_{jt} + \theta_k + e_{igjtk}$$

For student i, in grade g, in teacher j's classroom in year t in school k, her literacy achievement score ("$ElaAch_{igjtk}$") is a linear function of her prior achievement, a vector of student demographic characteristics including race, gender, English language status, and eligibility for free/reduced price lunch program (which serves as a proxy for socioeconomic status), grade level fixed effects, aggregated school level demographic characteristics, and a set of teacher-by-year fixed effects which indicates the teacher to which the student is exposed in the given year. The key parameter of interest is δ_{jt}, which captures the average achievement of teacher j's students in year t, conditional on prior skill and student characteristics, relative to the average teacher in the same grade.

The key change to this specification is the definition of "prior" test score timing. For the spring-spring data, I use the test score from the prior spring for $ElaAch_{igjt-1k}$. When I use the fall-spring data, I replace the prior year spring test score on the right-hand side of the equation with the fall score from the current year.

Findings / Results:
Description of the main findings with specific details.

Preliminary results suggest that growth during academic years (from fall to spring) is markedly steeper than growth during the summer periods (from spring to fall). (Insert Figure 1 about here). This figure demonstrates that learning appears to slow or even reverse directions during the summer period. The presentation will focus on quantifying the extent of this average effect as well as exploring the variance across teachers and students. Only to the extent that the summer learning experience varies by unobservable characteristics of the student will this finding bias value-added estimates.

Conclusions:
Description of conclusions, recommendations, and limitations based on findings.

Almost all value-added models estimate the effects of teacher using spring-to-spring outcome data, simply given that statewide annual testing occurs on this timeline. As a result, the teacher is attributed both with learning gains made by students during the school year as well as the prior summer. In essence, teachers are in part evaluated for what happens to students prior to their first day in their class, which runs counter to the fundamental goals of an accountability system. If this artifact of data collection introduces bias, it likely penalizes teachers whose students do not have positive learning during the previous summer, and it inappropriately rewards teachers whose students have a more enriched summer experience. Any accountability mechanism that hold teachers accountable for summers—a time period they cannot possibly influence—may have unintended consequences for the teacher workforce. This finding also highlights that the inferences that policy makers wish to make from statewide annual testing are not supported by the nature of the testing that is currently required to meet federal standards.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Alexander, Karl L.; Entwisle, Doris R.; Olson, Linda S. (2004). Schools, Achievement, and Inequality: A Seasonal Perspective. Summer learning: Research, policies, and programs. Borman, Geoffrey D. (Ed); Boulay, Matthew (Ed), Summer learning: Research, policies, and programs, (pp. 25-51). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, xii, 295 pp.

Alexander, Karl L.; Entwisle, Doris R.; Olson, Linda S. (2007). "Lasting Consequences of the Summer Learning Gap". *American Sociological Review*, vol. 72(2): 167-180

Cooper, H., B. Nye, K. Charlton, J. Lindsay, and S. Greathouse. (1996). "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review." *Review of Educational Research* 66: 227–68.

Downey, D. B., P. T. von Hippel, and B. Broh. 2004. "Are Schools the Great Equalizer? Cognitive Inequality During the Summer Months and the School Year." *American Sociological Review* 69:613–35.

Heyns, B. 1978. Summer Learning and the Effects of Schooling. New York: Academic.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.

McCaffrey, D. (2003). *Evaluating value-added models for teacher accountability*: Rand Corp.

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

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

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubin, D., Stuart, E., & Zanutto, E. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116.

Scherrer, J. (2011). Measuring Teaching Using Value-Added Modeling: The Imperfect Panacea. N*ASSP Bulletin* May 25, 2011.

Appendix B. Tables and Figures

Not included in page count.

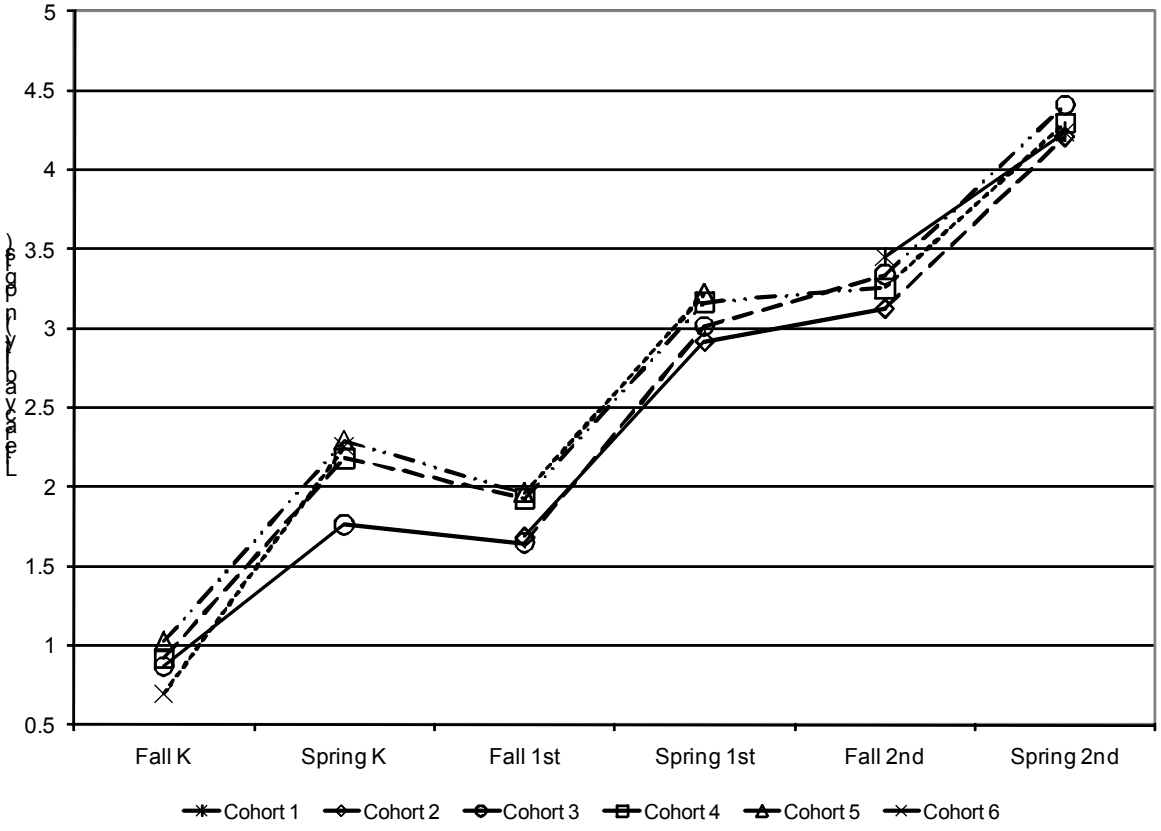

Figure 1. Means by cohort and year of Literacy Collaborative (LC) implementation.