Abstract Title Page

Title: Teacher Effects on Student Achievement and Height: A Cautionary Tale

Authors and Affiliations:

Marianne P. Bitler, Department of Economics, UC Irvine and NBER
Sean P. Corcoran, NYU Steinhardt School of Culture, Education, and Human Development
Thurston Domina, Department of Education, UC Irvine
Emily K. Penner, Department of Education, UC Irvine

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

The growing availability of data linking students to classroom teachers has made it possible to estimate the contribution teachers make to student achievement. By nearly all accounts, this contribution is large. Estimates of the impact of a one standard deviation (s.d.) increase in teacher "value-added" range from 0.10 to 0.30 s.d. in both math and reading. These effects have been documented in locations as diverse as Texas (Rivkin, Hanushek, & Kain, 2005), North Carolina (Clotfelter et al., 2006; Goldhaber & Hansen, 2012; Rothstein, 2010), Chicago (Aaronson, Barrow, & Sander, 2007), Florida (Harris & Sass, 2006; McCaffrey et al., 2009), New Jersey (Rockoff, 2004), San Diego (Koedel & Betts, 2009), Los Angeles (Buddin & Zamarro, 2008; Kane & Staiger, 2008), and elsewhere (Jacob, Lefgren, & Sims, 2010; Papay, 2011). The magnitude of these effects suggests a student assigned to an effective teacher will experience nearly a full year's more growth than a student assigned to an ineffective teacher.

Galvanized by these findings, policymakers at the state and school district level have moved to incorporate value-added measures into teacher evaluation systems. President Barack Obama and U.S. Secretary of Education Arne Duncan have embraced this approach and pressured states to use value-added measures as significant criteria in the promotion, compensation, and dismissal of educators. By some estimates, the potential for such policies to raise teacher quality, student learning, and economic growth is substantial (e.g., Chetty et al., 2011; Hanushek, 2009).

While there is a growing consensus that teacher quality is important and current evaluation systems are inadequate, many have expressed concerns over the use of value-added measures (VAMs) in high-stakes personnel decisions. These concerns are often grounded in VAMs' statistical imprecision and possible susceptibility to bias (Briggs & Dominique, 2011; Corcoran & Goldhaber, 2013; Harris, 2011). Because teachers are not randomly assigned to students, VAMs are plausibly biased by the presence of other unmeasured student, class, or school influences on achievement (Rothstein, 2010). Moreover, VAMs are imprecise, with a substantial proportion of the variation in achievement across classrooms attributable to student and classroom-level noise (McCaffrey et al., 2009; Schochet & Chiang, 2013). While research on these issues is ongoing, the prevailing view appears to hold that these limitations are not significant barriers to the use of VAMs in evaluating teacher performance, if done with appropriate caution (Glazerman et al., 2010, 2011).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

We conduct a new test of the validity of teacher value-added models. We apply traditionally-estimated VAM models to an outcome that teachers cannot plausibly have a causal effect on: student height. Any estimated "effect" of teachers on height should raise questions about the extent to which VAMs cleanly distinguish between effective and ineffective teachers. We also examine two potential interpretations for effects of teachers on height. The first is that these effects reflect bias, sorting to teachers on the basis of unobserved factors related to height (that

may or may not be related to achievement). The second is that these effects reflect measurement error or other forms of random "noise." Both have implications for the use of VAMs in practice.

Setting:
Description of the research location.

All students in grades 4 and 5 in New York City public schools between 2007 and 2010, and kindergarten students in the Early Childhood Longitudinal Survey – Kindergarten Cohort (ECLS-K), a national study of more than 20,000 children in the kindergarten class of 1998.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.

We use two data sources for estimating teacher effects on achievement and height. The first is a panel of all students in grades 4-5 in New York City public schools between 2007 and 2010 (approximately 473,000 student-year observations). Each student is linked to their English Language Arts (ELA) and mathematics teacher, and to annual "Fitnessgram" measurements of their height. The second is the Early Childhood Longitudinal Survey – Kindergarten Cohort (ECLS-K). ECLS-K students are linked to classroom teachers, and were tested in reading and math at the beginning and end of the school year. Trained assessors additionally measured participants' height in both the fall and spring of the kindergarten year. We use approximately 9,200 students from the ECLS-K who shared a classroom teacher with at least 4 other students.

These two data sources contribute to our analysis in different ways. The NYC data represents a large population of students and teachers over four years. The number of students observed per teacher is relatively large, allowing for precise estimates of teacher effects and estimation across multiple years of data. The ECLS-K sample is smaller, but height measurements in this sample are less likely to be measured with error. Another advantage of the ECLS-K is that there will be less sorting within schools to teachers on the basis of unobserved factors related to achievement (or height). To the extent non-random sorting of students is a problem in typical value-added models (e.g., Rothstein, 2010), this should be less of a concern in kindergarten (and thus in the ECLS-K), where teachers should know less about incoming students. The ECLS-K achievement scores are continuous, normally distributed, and devoid of ceiling effects, while NYC scores have a lumpier distribution. The ECLS-K provides us a national sample. Finally, the ECLS-K conducted extensive teacher and parent interviews, which allow us to control for more family background characteristics than are typically available in administrative data sets.

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.

In the traditional estimation of teacher value-added on student achievement, individual teachers are treated as non-experimental "interventions" for which impacts can be estimated. Value-added models assume that systematic variation in achievement across teachers can provide an unbiased estimate of teachers' causal effects on test performance, after controlling for prior achievement and a limited set of covariates. In our NYC analysis, 4^{th} and 5^{th} grade teachers are the "intervention" of interest. In the ECLS-K, kindergarten teachers are the intervention of interest.

Research Design:
Description of the research design.

For each NYC outcome (math and ELA achievement, height) we estimate teacher effects using a standard value-added model that controls for the prior year outcome, year indicators, and student covariates (gender, race/ethnicity, recent immigrant and LEP status, special education status, eligibility for free or reduced-price lunch, and borough). Models for height also include age-by-gender interactions and race-by-gender interactions. Teacher effects are alternately estimated assuming random or fixed effects. The random effects are the Empirical Bayes, or best linear unbiased predictors (BLUPs) of the teacher effects; we adjust the fixed effects to account for sampling variation. In practice, the random and fixed effect estimates are very similar.

To the extent teacher effects are biased due to school-level influences on the outcome of interest correlated with teacher assignment (e.g., leadership quality or other resources), we also estimate teacher effect models with school effects included. This is done in two steps, first regressing current achievement on all covariates and school effects, and then estimating teacher effects from the residuals. (We note that while the estimation of value-added models with school effects is common in the research literature, it is deliberately not done often in practice).

For ECLS-K outcomes we estimate a model similar to that for NYC, but using spring and fall measures as the current and lagged dependent variables, respectively. Covariates include gender, race/ethnicity, and an indicator for whether the child speaks a language other than English at home. Teacher effects are again estimated assuming random or fixed effects, and we account for the complex sampling design of ECLS-K by using the kindergarten panel weight, and in the random effects models, the school weight as well.

We begin by summarizing the magnitude of the estimated teacher effects on each outcome, measured by the s.d. of the random or fixed effect. These are compared across outcomes (e.g., height and achievement) and model specifications (e.g., models with school effects and without). We then examine how teachers' effects are correlated across outcomes. A correlation between value-added in height and in achievement could be indicative of sorting to teachers on the basis of unobserved factors related to each. We compute within-teacher correlations in effects across years as a measure of the persistent component of effects across years. Finally, as an alternative test for the role of noise, we re-estimate each model after randomly allocating observed student data across teachers and schools. This approach eliminates all possibilities of non-random sorting of students to teachers and thus remaining "effects" should be pure noise.

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

All NYC data comes from administrative databases provided by the New York City Department of Education. Restricted use ECLS-K data was obtained from the National Center for Education Statistics under a licensing agreement that ensures data anonymity and security. Our procedure for analyzing the data was summarized in the previous section.

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

In both data sources we find the magnitude of teacher effects on height is nearly as large as their effect on math and reading achievement. For instance, we find a one s.d. increase in teacher "value-added" to the height of New York City 4th graders is associated with 0.23 s.d. taller students. This effect size can be compared to a 0.28 and 0.25 s.d. impact on achievement in math and ELA, respectively. Models that include school effects reduce the magnitude of these estimates, although they remain large at 0.16-0.17 s.d. We find very similar effects of teachers on height in the ECLS-K. We find no direct evidence of bias in achievement VAMs, given we find a correlation between teachers' value-added in height and achievement that is zero. By the same token, we observe no correlation in teacher effects on height across years, an indication that there is no persistent component to teacher effects on height. Nevertheless, the extent of noise in these models appears to be large. Even when randomly assigning student data to teachers and schools, we continue to find a significant "effect" of teachers on achievement.

Conclusions:

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

These findings raise important questions about the extent to which VAMs cleanly distinguish between effective and ineffective teachers. Our finding of a near-zero correlation between a teacher's value-added on height and her value-added on achievement offers some comfort that VAMs for achievement are not biased by the presence of omitted variables associated with height. Moreover, our finding of a near-zero correlation in a teacher's effect on height across years suggests that the "signal" contained in these effects is negligible. However, our results do show that the "noise" in value-added models is substantial. When fitting traditionally-estimated value-added models to an outcome in which teachers cannot plausibly have an impact, we find "effects" that are as large as those found for outcomes that teachers do affect.

Taken together, our results provide a cautionary tale for the interpretation and use of teacher VAM estimates in practice. We find that—simply due to chance—teacher effects can appear large, even on outcomes they cannot plausibly affect. The implication is that many value-added studies likely overstate the extent to which teachers differ in their effectiveness, although further research is needed. Furthermore, users of VAMs should take care to ensure their estimates reflect the signal component of teacher effectiveness and are not driven by noise. This is especially important when personnel and compensation decisions are tied to individual VAM estimates. While most contemporary value-added systems do adjust VAMs for noise (applying a "shrinkage" factor), the procedure for doing so is not standard, and the quality of these adjustments is unclear.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Briggs, D., & Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the "Los Angeles Times." *National Education Policy Center*.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics*, *66*(2), 103–115.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. NBER Working Paper No. 17699.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Corcoran, S., & Goldhaber, D. (2013). Value added and its uses: Where you stand depends on where you sit. *Education Finance and Policy*, *8*, 418–434.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems.* Washington D.C.: Brookings Institution.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Washington D.C.: Brookings Institution.

Goldhaber, D., & Hansen, M. (2012). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*, 589–612.

Hanushek, E. A. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165–80). Washington D.C.: The Urban Institute Press.

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

Harris, D. N., & Sass, T. R. (2006). *Value-added models and the measurement of teacher quality*. Working Paper, Florida State University.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, *45*(4), 915–943.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Working Paper No. 14607, National Bureau of Economic Research.

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

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

Papay, J. P. (2011). Different tests, different answers The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163–193.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

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*(2), 142–171.

Appendix B. Tables and Figures

Not included in page count.