ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-172

Teacher Effects on Student Achievement and Height: A Cautionary Tale

Marianne Bitler
UC Davis, and NBER

Sean Corcoran
Vanderbilt University

Thurston Domina
University of North Carolina

Emily Penner
UC Irvine

Estimates of teacher "value-added" suggest teachers vary substantially in their ability to promote student learning. Prompted by this finding, many states and school districts have adopted value-added measures as indicators of teacher job performance. In this paper, we conduct a new test of the validity of value-added models. Using administrative student data from New York City, we apply commonly estimated value-added models to an outcome teachers cannot plausibly affect: student height. We find the standard deviation of teacher effects on height is nearly as large as that for math and reading achievement, raising obvious questions about validity. Subsequent analysis finds these "effects" are largely spurious variation (noise), rather than bias resulting from sorting on unobserved factors related to achievement. Given the difficulty of differentiating signal from noise in real-world teacher effect estimates, this paper serves as a cautionary tale for their use in practice.

VERSION: December 2019

Suggested citation: Bitler, Marianne, Sean Corcoran, Thurston Domina, and Emily Penner. (2019). Teacher Effects on Student Achievement and Height: A Cautionary Tale. (EdWorkingPaper: 19-172). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/k02z-9n94

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

Marianne Bitler, Professor of Economics, UC Davis, and NBER;

Sean Corcoran, Associate Professor of Public Policy and Education, Peabody College of Education and Human Development, Vanderbilt University;

Thurston Domina, Professor of Education, University of North Carolina;

&

Emily Penner, Assistant Professor of Education, UC Irvine

Abstract

Estimates of teacher "value-added" suggest teachers vary substantially in their ability to promote student learning. Prompted by this finding, many states and school districts have adopted value-added measures as indicators of teacher job performance. In this paper, we conduct a new test of the validity of value-added models. Using administrative student data from New York City, we apply commonly estimated value-added models to an outcome teachers cannot plausibly affect: student height. We find the standard deviation of teacher effects on height is nearly as large as that for math and reading achievement, raising obvious questions about validity. Subsequent analysis finds these "effects" are largely spurious variation (noise), rather than bias resulting from sorting on unobserved factors related to achievement. Given the difficulty of differentiating signal from noise in real-world teacher effect estimates, this paper serves as a cautionary tale for their use in practice.

We thank the NYC Department of Education and Michelle Costa for providing data. Amy Ellen Schwartz was instrumental in lending access to the Fitnessgram. Greg Duncan, Avi Feller, Richard Startz, Jim Wyckoff, Dean Jolliffe, Richard Buddin, George Farkas, Sean Reardon, Michal Kurlaender, Marianne Page, Susanna Loeb, Jesse Rothstein, Jeff Smith, Howard Bloom, and seminar participants at Teachers College Columbia University, Stanford CEPA, Irvine Network on Interventions in Development, and APPAM provided helpful comments. Siddhartha Aneja, Annie Laurie Hines, and Danea Horn provided outstanding research assistance. All remaining errors are our own. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P01HD065704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1 Introduction

The increased availability of data linking students to 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 (σ) increase in teacher "value-added" on math and reading achievement typically range from 0.10 to 0.30σ, which suggest that a student assigned to a more effective teacher will experience nearly a year's more learning than a student assigned to an less effective teacher (Hanushek & Rivkin 2010; Harris 2011; Jackson, Rockoff, & Staiger 2013; Koedel, Mihaly, & Rockoff 2015). These estimates—and evidence that teacher value-added to student achievement is predictive of long-run outcomes (Chetty et al. 2014b)—provide the basis for the oft-cited assertion that teachers are the most important school input into student learning.

Prompted by these findings, policymakers have moved to adopt value-added measures (VAMs) as significant criteria in the evaluation, promotion, compensation, and dismissal of teachers. As of the 2015-16 school year, 36 of 46 states with revised teacher evaluation systems had incorporated VAMs or comparable student achievement growth measures into teachers' annual evaluations (Steinberg & Donaldson 2016). If such policies successfully improve teacher quality (Hanushek 2009), they have the potential to have a large impact on economic growth over the long run (Chetty et al. 2014b).

Despite the established consensus that teacher quality is important, concerns have been raised over the validity and reliability of VAMs for high-stakes personnel decisions (e.g., Baker et al. 2010; Braun, Chudowsky, & Koenig 2010). Because teachers are not randomly assigned to students, VAMs are potentially biased by student, classroom, or school influences on achievement that vary with teacher assignment (Dieterle et al. 2014; Kane 2017; Horvath 2015; Porvath & Amerein-Beardsley 2014; Rothstein 2010). Even if VAMs are not biased, they may lack stability if a large share of their variability is attributable to student and classroom-level variation unrelated to teachers (McCaffrey et al. 2009; Schochet & Chiang 2013). These concerns notwithstanding, the prevailing view appears to be that VAMs, though lacking in causal interpretation, are useful for classifying and evaluating teachers (Glazerman et al. 2010, 2011; Kane et al. 2013; Koedel, Mihaly, & Rockoff 2015; Sass, Semykina, & Harris 2014).

In this paper, we provide a stark illustration of the limitations to using value-added models to identify high- and low-performing teachers. We do this by applying commonly estimated models to an outcome that teachers cannot plausibly affect: student height. Aside from the implausibility of teacher effects on height, student height is an attractive measure for this exercise since it is symmetrically distributed, interval measured, and arguably less prone to measurement error than achievement. We find that the estimated teacher "effects" on height are nearly as large as the variation in teacher effects on math and reading achievement. Using a common measure of effect size in standard deviation units, we find a 1σ increase in "value-added" on the height of New York City 4th graders is about 0.22σ, or 0.65 inches. This compares to 0.29σ and 0.26σ, in math and English language arts, respectively. Moreover, this variation is statistically significant when measured using permutation tests. Models that control for school effects reduce the dispersion in effects on height, although the effects remain large and comparable to those on achievement at $0.16\sigma - 0.17\sigma$.

On their face, findings of teacher effects on height raise concerns about what these models are measuring. We consider three possible interpretations. The first, and potentially the most worrying, is that they reflect sorting to teachers on unobserved factors related to height that are also related to achievement. Under this interpretation, achievement differences attributed to teachers could reflect an unobserved sorting process, resulting in biased VAM estimates. The second interpretation is that teacher effects on height reflect spurious or sampling variation, or "noise." In this scenario, differences attributed to teachers are simply idiosyncratic variation across years and relatively small samples of students. Finally, there could be sorting on unobserved factors related to height that are uncorrelated with achievement. This type of sorting would be less worrisome for use in educational settings, as it would not imply systematic bias in estimates of teacher effects on test performance.

We evaluate these explanations in several ways. First, we examine the correlation between teachers' estimated effects on height and achievement. If effects on height reflect sorting on unobserved factors related to achievement, one might expect these effects to be correlated. Instead, we find a correlation close to zero. Second, we use a method proposed by Horvath (2015) to identify schools that

appear to systematically sort students to classrooms on the basis of prior characteristics. While more than 60% of NYC schools appear to track students on prior achievement, we find little evidence of tracking on height. Third, we estimate the "persistent" component of teacher effects using the covariance across years, for teachers with multiple years of data. While positive for achievement measures, the covariance for height is close to zero, suggesting the observed effect of teachers on height is largely spurious variation. Finally, we perform a series of permutation tests that randomly allocate students to teachers in our data set and re-estimate each VAM model. This approach eliminates any potential for sorting, peer effects, systematic measurement error (e.g., at the classroom or school level), and/or true effects; and provides a benchmark for what teacher "effects" look like simply due to noise or sampling variation. Using this benchmark, we can reject the null hypothesis of a zero standard deviation in teacher effects on height, suggesting the presence of at least some systematic unexplained variation across teachers.

Taken together, our results provide a cautionary tale for the use and interpretation of value-added models as they are often used in practice. We show that—simply due to chance—teacher effects can appear quite large, even on outcomes teachers cannot plausibly affect. To ameliorate the effect of sampling error on value-added estimates, analysts often apply a "shrinkage" factor, which scales VAMs by their estimated signal-to-noise ratio (Herrmann et al. 2016). This procedure is not always done, nor is it done in the same manner. In our context, the only approach to shrinkage that provides the theoretically "correct" adjustment—that is, shrinkage of height effects to the mean of zero—uses the covariance in effects from multiple years of classroom data to estimate the signal component (as in Kane & Staiger 2008, and Kane, Rockoff, & Staiger 2008). Shrinkage approaches that do not make use of classroom-level variation within teachers over time to estimate signal continue to yield nonzero estimates of teacher effects on height. At the same time, we argue that getting the shrinkage factor "right" may be of limited value for adjusting teacher effects in practice, since shrinkage has only modest effects on the relative *rankings* of teachers, which are typically what teacher evaluation systems rely upon.[1]

[1] On this point, see also Guarino et al. (2015) and Hermann et al. (2016).

In the next section, we provide the framework for our analysis and ground our work in the context of a large literature on teacher value-added models. Then, in Sections 3 and 4 we describe our data sources and empirical approach. Section 5 presents our main results and a set of robustness checks, and Section 6 concludes with a discussion and lessons for researchers and policymakers.

2 Background: estimation and properties of VAMs

Teacher effects are defined as the systematic variation in student test performance across teachers that remains after accounting for the effects of other observed inputs, such as prior academic achievement, and economic or educational disadvantages. They are generally estimated from a model like the following:

$$Y_{it} = \alpha Y_{it-1} + X'_{it}\beta + \epsilon_{it} \quad (1)$$

in which Y_{it} and $Y_{it-1}$are test scores for student i in years t and $t-1$, respectively, and X_{it} is a vector of student-level covariates related to achievement (and potentially, teacher assignment). ϵ_{it} is an error term that can be decomposed into teacher (j) and student-level errors, or when multiple years of classroom data are available, teacher, classroom, and student errors:

$$\epsilon_{it} = u_j + e_{it}, \qquad or \quad \epsilon_{it} = u_j + v_{jt} + e_{it} \quad (2)$$

Estimates of σ_u—the standard deviation in teacher effects on student achievement—typically range from 0.10σ, to 0.30σ, with effects typically larger in mathematics than in reading.[2] Especially in comparison with other educational interventions, the impact of differences in teacher quality appears to be substantial.

The residual variation (ϵ_{it}) used to estimate teacher effects consists of a "true" (persistent) teacher effect, effects of any unmeasured factors related to achievement that are common to a teacher, and idiosyncratic student- and classroom-level errors (Ishii & Rivkin 2011; Koedel, Mihaly, & Rockoff

[2] For extensive reviews of this literature, see Hanushek & Rivkin (2010), Harris (2011), Jackson, Rockoff, & Staiger (2013), and Koedel, Mihaly, & Rockoff (2015). As described later, some of these variance estimates are adjusted for sampling error, while others are not.

2015).[3] If not accounted for, the latter factors have the potential to inflate estimates of the overall impact of teacher quality (i.e., estimates of σ_u), or—more importantly for practice—bias individual estimates of teacher effects.

2.1 Unobserved sorting and the question of bias

A large literature has investigated whether and to what extent VAMs are systematically biased. For example, a number of studies have asked whether model specification—the inclusion or exclusion of student or classroom controls, for example—affects VAM estimates (e.g., Ballou, Sanders, & Wright 2004; Ballou, Mokher, & Cavalluzzo 2012; Ehlert et al. 2013; Goldhaber, Goldschmidt, & Tseng 2013; Sass, Semykina, & Harris 2014; Kane et al. 2013). With the exception of a control for prior achievement, these studies tend to find that the choice of covariates has only modest effects on the relative rankings of teachers. Rankings tend to be more sensitive to the inclusion of school fixed effects, which allow for systematic variation in achievement across schools due to sorting or other school-level inputs. However, because school effects absorb real differences in mean teacher effectiveness across schools, they are more commonly used in research than in practical applications such as district level estimates of teacher value added (Ehlert et al. 2014; Goldhaber, Walch, & Gabele 2013; Gordon, Kane, & Staiger 2006; Kane & Staiger 2008).

A more worrisome concern is that students are assigned to teachers on the basis of time-varying factors observed to schools but unobserved by the analyst. In a notable test of bias of this type, Rothstein (2010) showed that teachers assigned to students in the future had statistically significant "effects" on contemporaneous achievement gains. Because such effects cannot be causal, Rothstein argued that VAMs inadequately account for the process by which students are assigned to teachers.[4] In another study, Kane and Staiger (2008) randomly assigned teachers to classrooms within Los Angeles schools and found non-experimental VAM estimates were generally unbiased predictors of experimental VAMs, suggesting little

[3] Some value-added models allow for "drift" in teacher effectiveness over time; for example, see Chetty et al. 2014a.
[4] Several subsequent papers have argued the "Rothstein test" may not be robust (Goldhaber & Chaplin 2012; Kinsler 2012; Koedel & Betts 2011).

bias (at least within schools). This finding was replicated in the larger Measures of Effective Teaching (MET) project (Kane et al. 2013), and a quasi-experimental study by Chetty et al. (2014a) that focused on teachers switching between schools found little evidence of bias (see also Bacher-Hicks, Kane, & Staiger 2014; Bacher-Hicks et al. 2017).[5]

2.2 Statistical imprecision, instability, and noise

Even if VAMs are unbiased, their utility in evaluating individual teachers' job performance could be limited by statistical imprecision and instability (McCaffrey et al. 2009; Schochet & Chiang 2013). Imprecision stems not only from sampling error—a consequence of the small number of students used to estimate teacher effects—but also from classroom-level shocks, and poor model fit, particularly for teachers with students in the tails of the distribution or with otherwise hard-to-predict achievement (Herrmann et al. 2016; Kane 2017). McCaffrey et al. (2009) found that 30 to 60 percent of the annual variation in teacher effect estimates from Florida were a result of random variation at the student level; of the remainder, 50 to 70 percent could be considered "persistent" teacher effects and the rest random classroom-level variability. A practical implication of imprecision is that annual VAM estimates vary from year to year, sometimes substantially, with correlations ranging from 0.18 to 0.64, raising the possibility that seemingly effective teachers in one year are judged ineffective in the next, and vice versa.[6] Studies that report cross-year correlations include, for example, Aaronson, Barrow, & Sander (2007), Chetty et al. (2014a), and Goldhaber & Hansen (2013). Stability depends a great deal on model specification; for example, whether student or school fixed effects are used (Koedel, Mihaly, & Rockoff 2015). A counter to this concern is that value-added measures—despite their instability—are useful predictors of career productivity, and better than alternative measures of teaching effectiveness (Glazerman et al. 2011; Goldhaber & Hansen 2013; Staiger & Kane 2014).

[5] Another potential source of bias is through test scaling. See, for example, Kane (2017), Soland (2017), and Briggs & Dominigue (2013).

[6] Studies that report cross-year correlations include, for example, Aaronson, Barrow, & Sander (2007), Chetty et al. (2014a), and Goldhaber & Hansen (2013). Stability depends a great deal on model specification, for example, whether student or school fixed effects are used (Koedel, Mihaly, & Rockoff 2015).

A common procedure used to address imprecision in value-added estimates is Empirical Bayes shrinkage, which scales unadjusted estimates by a shrinkage factor λ_j, the signal-to-noise ratio (Guarino et al. 2015; Hermann et al. 2016; Kane, Rockoff, & Staiger 2008; Koedel, Mihaly, & Rockoff 2015):

$$\lambda_j = \frac{\sigma_u^2}{\sigma_u^2 + \left(\sigma_e^2 / n_j\right)} \tag{3}$$

λ_j depends on n_j, the number of student observations for teacher j, and the estimated proportion of the overall variation in achievement that is attributable to teacher effectiveness. (σ_u^2 is the between-teacher variance in student achievement, and σ_e^2 is the within-teacher variance). Intuitively, a teacher effect estimate is "shrunk" toward the mean of zero if it is estimated using a small number of students, or if the share of the overall variation in estimated teacher effects that is "signal" versus "noise" is low. Thus, individual teacher effects estimated off a relatively small number of students—as well as individual teacher effects that vary dramatically from the mean teacher effect (here set to 0)—are disproportionately "shrunk" toward zero. Overall, a teacher effect estimate is "shrunk" toward zero the smaller is: (1) the number of students used to estimate the effect, and (2) the share of the overall variation in estimated teacher effects that is "signal" versus "noise."

Shrinkage requires having appropriate estimates of the signal (σ_u^2) and noise (σ_e^2) variance components (Schochet & Chiang 2013). Studies vary in the extent to which they use shrinkage at all, and in their method for estimating σ_u^2 and σ_e^2 (Guarino et al. 2015). One approach is to obtain the best linear unbiased predictors (BLUPs) of u_j from a teacher random effects model, which are by definition Empirical Bayes shrinkage estimates. Another used by Kane, Rockoff, and Staiger (2008) uses teachers with multiple years of classroom data to calculate the covariance in classroom effects to provide an estimate of σ_u^2. This method begins by estimating teacher-by-year effects u_{jt} (equal to $u_j + v_{ijt}$ in Equation 1). Under the assumption that teacher effects are constant over time, the covariance between classroom effects for the same teacher across years is equal to the teacher effect variance σ_u^2. Chetty et al. (2014a) relax the assumption of constant effects over time and allow for drift, but the idea is the same.

Other authors allow for heteroskedasticity in the within-teacher (classroom) error variance (e.g., Herrmann et al. 2016).

The shrinkage approach is limited in several respects, however. First, multiple years of classroom data are not always available to the analyst or used. For example, many teacher evaluation systems calculate VAMs from a single year of achievement data (see American Institutes for Research 2013; Isenberg & Hock 2010; VARC 2010). Second, as we discuss in a later section, shrinkage has only a modest impact on the relative ranking of teachers when there is little variation in the number of students per teacher (see also Guarino et al. 2015; Herrmann et al. 2016). In the extreme case when n_j is the same for all teachers, shrinkage has no effect on relative rankings. A teacher evaluation system that rewarded or punished teachers based on relative effects would make the same decisions regardless of what shrinkage factor was used. In sum, shrinkage procedures are at best a partial solution to the presence of non-persistent variation in student outcomes across teachers.

In the next section, we describe the data used in our estimates of teacher value-added on achievement and height.

3 Data

Our primary data source for estimating teacher effects on height and achievement is a panel of more than 360,000 students enrolled in grades 4-5 in New York City public schools between 2007 and 2010. These data are well-suited to our purposes, for several reasons. First, each student is linked to their mathematics and English Language Arts (ELA) teacher and to annual measurements of their height from the city's "Fitnessgram" physical fitness assessment. Second, the data represent a large population of students and teachers over four years. The number of students observed per teacher is large for some teachers, allowing for more precise estimates of teacher effects and estimates using multiple years of classroom data. Third, these data are typical of those used to estimate teacher VAMs in practice and were in fact used by the NYC Department of Education to evaluate teachers' effectiveness in math and ELA

(Rockoff et al., 2012).[7] The Fitnessgram also includes measures of student weight. We do not report results using weight in the interest of brevity, and because teachers may have real "effects" on weight (e.g., through their practices related to physical activity, such as recess participation and school meals/snacks).

We began with an administrative panel data set for students enrolled in grades 3-5 between 2005-06 and 2009-10. Among other things, this panel included student demographics (birth date, gender, race/ethnicity), program participation (ELL, special education, and participation in the free and reduced school meals program (a measure of eligibility among those who apply for the program), and scale scores in math and ELA, which we standardized by subject, grade, and year to mean zero and standard deviation one. The administrative data were matched to teacher-student linkages in math and ELA from 2006-07 to 2009-10. Linkages were also available for 2010-11, but teacher codes changed in that year as a result of the NYCDOE's switch to a new personnel system. This change prevented us from matching teachers in 2010-11 to earlier years. Although students in grades 6-8 could also be linked to teachers, we restricted our analysis to elementary school students, who are predominately in self-contained classrooms with one teacher for core subjects. This approach allowed us to avoid issues of proper attribution to middle school teachers. Third grade records and 2005-06 data were retained only to provide lagged values of the outcome measures.

The Fitnessgram has been conducted annually in NYC public schools since 2005-06 and relies on school staff—usually the physical education teacher—to measure students' height, weight, and physical fitness. School personnel are trained to collect height and weight using a common procedure and a recommended digital beam scale. See https://vimeo.com/album/4271100/video/217670950. Measurements are taken throughout the school year, and the date of measurement is recorded in the Fitnessgram data. To parallel the measures used in our achievement models, we standardized height in

[7] Bitler et al. (2015) perform a similar analysis using the nationally representative Early Childhood Longitudinal Survey-Kindergarten Cohort (ECLS-K), which links students to classroom teachers and also includes measures of achievement and health. Their findings are similar.

inches by grade and year to mean zero and standard deviation one, with outlying values more than 4σ from the mean set to missing before standardizing. As explained later, we experimented with other methods for standardizing height, such as by gender and age in months. The reference group for standardization had little to no effect on our results. In all cases, we standardized using all available data, not the analytic sample, which was more restrictive.

Descriptive statistics for students in our analytic sample are reported in Table 1, alongside statistics for the full population of students who could be linked to classroom teachers. Students in the analytic samples for height, math, and ELA were required to have a non-missing lagged dependent variable, non-missing covariates, and a teacher with seven or more students in the same grade with enough data to be included in the VAM models. Seven is a common minimum group size requirement used in other studies and in state teacher evaluation systems. For our baseline models which combine all four years of data, this minimum group size of 7 is not that restrictive. Table 1 shows the average 4th and 5th grader in our analytic sample was somewhat higher-achieving and smaller in stature than the full population of students linked to classroom teachers, with marginally higher ELA and math scores. (This is common for the analytic sample used to estimate value-added models, since students are required to have a lagged test score). The average 4th and 5th grader in NYC was 54.7 and 57.1 inches in height, respectively, with standard deviations of 3.0 and 3.2 inches. For later reference, the average 5th grader grew 2.5 inches between 4th and 5th grade, with a standard deviation of 1.8.

Table 2 reports the number of unique teachers and classrooms in our analytic samples, and descriptive statistics for the number of students per teacher (pooling all years) and per class, or teacher-year. The full distributions are shown in supplemental appendix Figures A.1 and A.2. Over the four years combined, approximately 4,300 to 4,700 4th grade teachers and 3,700 to 4,200 5th grade teachers were in the analytic sample, depending on the outcome measure. There were more teachers in the analytic sample for math, as students were more likely to be missing data for height or ELA. Some teachers were observed with 80 or more students over four years, although the average teacher was observed for only two years, with 36 to 42 students. The average number of students per classroom (teacher-year) was 20-

21 for all outcomes. Teachers in our math sample represent about 82-84 percent of all 4th-5th grade NYC teachers who could be linked to students during these years. Similarly, the teachers in our height and ELA samples represent 73-74 and 80-82 percent of all grade level teachers, respectively.

To get a better sense of the underlying scales of our outcome variables, Figures 1 and 2 show histograms of student height and math achievement in the 4th and 5th grade analytic samples. For height, we show both the original measure in inches and the standardized measure. The distributions of both measures are roughly bell-shaped, although not normal: K-S tests reject normality, and there are a few low-scoring outliers in math. A small mass of students also scored at or near the test ceiling in math. The original height distribution has a relatively small number of discrete values (26 and 27 unique values in 4th and 5th grades, respectively), although after standardization the measure takes on a larger number of values, since standardization is within grade and year. (The number of discrete scale scores in math were 65 and 43 in 4th and 5th grades, respectively). The discrete nature of the height measure in NYC could potentially affect the dispersion of teacher effects.

Finally, Table 3 reports student-level pairwise correlations between the height, math, and ELA measures, between year-to-year changes in these measures, and between each measure and its lag. While the z-scores for math and ELA are strongly correlated (0.69 and 0.59 in 4th and 5th grade), achievement has only a weak bivariate correlation with height. The small negative correlation could be due to grade repeaters, who would be tall for their grade. In a multivariate regression model for achievement that includes height as a predictor (with controls for lagged achievement, age, and other standard covariates), height is a statistically significant predictor of achievement in math and ELA for both grades 4 and 5. The implied effect size is small, however, with a 1σ increase in height associated with a 0.011σ to 0.015σ higher test score. See supplemental appendix Table A.1 for details. All three measures are strongly correlated with their lagged values, with correlations ranging from 0.65-0.68 in ELA and math to 0.79-0.80 in height. The strong year-to-year correlation in height suggests it is reliable. The correlations between year-to-year changes in height and year-to-year changes in achievement are very low.

4 Empirical methods

4.1 Baseline value-added models

For each grade level and outcome (math achievement, ELA achievement, and height) we estimated teacher effects using a standard "dynamic OLS" value-added model that conditions on the prior year's outcome and a set of student-level covariates:

$$Y_{it} = \alpha Y_{it-1} + X'_{it}\beta + \gamma_t + u_j + e_{it} \quad (4)$$

y_{it} and y_{it-1} are outcomes for student i in years t and $t-1$, respectively, and X_{it} is a vector of fixed and time-varying characteristics of student i. The covariates in X_{it} include a three-way interaction of gender, race, and age; recent immigrant status; limited English proficiency (LEP) and an indicator for a language other than English spoken at home; special education status; participation in free or reduced-price lunch (a measure of eligibility among those who apply for the program); and borough of residence. Free or reduced-price lunch indicators are missing for some students, typically those enrolled in universal free meals schools, where schools provide free meals to all students regardless of income eligibility. We coded these students with a zero but included an indicator equal to one for students with missing values. The three-way interaction of gender, race, and age is not standard in VAM models, but was thought to be more appropriate in the model for height. Results are nearly identical with non-interacted controls. In their review of the literature; Koedel, Mihaly, and Rockoff (2015) note that after conditioning on lagged outcomes, value-added estimates are not especially sensitive to the choice of covariates. Like other authors, they also caution against estimating teacher effects on test score gains. Since the gains model is seldom used in recent work, we do not take that approach here. The covariates in the height model were identical to those used for math and ELA, with the exception of an additional control for days elapsed between the student's annual Fitnessgram measurements across years, the timing of which can vary between and within schools. The γ_t are year effects, and the u_j are the teacher effects of interest.

To capture the variety of ways in which teacher effects are estimated in practice, we estimate the u_j alternately under fixed and random effects assumptions.[8] As described in Koedel, Mihaly, & Rockoff (2015); there are advantages and disadvantages to each approach, and recent work has converged on the fixed effects specification. We estimated random teacher effects in two ways. The first fit a random effects model using maximum likelihood and then obtained the best linear unbiased predictor (BLUP) of each teacher effect. The second estimated Equation 2 without teacher effects and then obtained the mean residual for each teacher j. The mean residuals for teacher j were then manually "shrunk" as described in Section 4.2. Fixed effects were estimated using OLS and multiplied by a shrinkage factor for comparability with the random effects. We refer to these as "adjusted" fixed effects. Left unadjusted, their standard deviation would overestimate the variability in teacher effects. As is commonly found in other studies, we find VAMs estimated under random and fixed assumptions are highly correlated (0.71 to 0.96, depending on the grade and outcome measure, and in most cases above 0.90).

We also estimated versions of Equation (4) with school fixed effects ϕ_s:

$$Y_{it} = \alpha Y_{it-1} + X'_{it}\beta + \gamma_t + \phi_s + u_j + e_{it} \quad (5)$$

As noted in Section 2, value-added models with school effects are less common in practical applications than in research, and have a number of disadvantages (Koedel, Mihaly, & Rockoff 2015). In our context, however, we were concerned that height measurement practices could vary at the school level (e.g., measurement tool or assessor-level variation). School-level sorting by factors correlated with height is also a possibility. The school effects ϕ_s will absorb time-invariant factors of these types (but not time-varying aspects). In the achievement models, the school effects should capture the effects of time-invariant school influences on the outcome, including leadership quality, other staff, special academic programs, or other resources.

Equation 3 was estimated in two steps (as in Master, Loeb, & Wyckoff 2017). First, we regressed the outcome Y_{it} on all of the regressors in 3, including school effects, but excluding teacher effects u_j.

[8] The baseline models for value added are reported in supplemental appendix Tables A.2-A.4.

Residuals from this first step were then used in the second to estimate the teacher effects as either random or fixed effects, in the manner described above.

4.2 Variance components and classroom effects

The approach to shrinkage described in Section 4.1 uses two variance components: variance between and within teachers. Estimates of σ_u^2 and σ_e^2, together with the number of observations per teacher n_j, provide the factor λ_j in Equation 3. The goal of shrinkage in practice is to adjust value-added estimates for sampling error, recognizing that teacher effects are less precise when the number of students used to estimate each teacher effect is small. In our setting, the n_j over four years of data is large for some teachers, but smaller for others.

In the language of multilevel models, Equation 2 is a "2-level" model with students nested within teachers. The 2-level model does not, however, allow for random variation at the classroom level. Teachers in a 2-level model—even when observed with many classrooms—are observationally equivalent to a large classroom. A 3-level model allowing for a random classroom effect $\left(u_{jt} = u_j + v_{jt}\right)$ introduces unobserved, group-level variability within teachers over time (McCaffrey et al., 2009). In the 3-level model, the shrinkage factor includes a classroom variance component σ_v^2:

$$\lambda_j = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \left(\sigma_e^2 / n_j\right)} \tag{6}$$

σ_u^2, σ_v^2, and σ_e^2 can be estimated directly in the maximum likelihood procedure. Alternatively, the approach used by Kane, Staiger, and Rockoff (2008) estimates between-teacher variance (σ_u^2) as the covariance between classroom effects $\widehat{u_{jt}}$ for the same teacher in successive years, for teachers with multiple years of data. The between-classroom variance (σ_v^2) is estimated using the total variance in the residuals less the within-classroom and between-teacher components.

We used estimated teacher-by-year (classroom) effects $\widehat{u_{jt}}$ to test for correlation in teacher effects across years, and to obtain teacher effect estimates that could be shrunk using the Kane et al.

approach and equation 6). We also estimated the 3-level random effects model directly using maximum likelihood.

4.1 Testing for sorting on prior characteristics

To assess the extent to which students are non-randomly sorted to classrooms within schools on prior characteristics, including height, we used a method developed by Horvath (2015). For related approaches, see Dieterle et al. 2015; Aaronson, Barrow, & Sander 2007; and Clotfelter, Ladd, & Vigdor 2006. In that study, Horvath identified schools with non-random sorting by testing for systematic variation in lagged student characteristics across classrooms within schools, grades, and years. For example, for each school *s* the following regression is estimated for the lagged outcome y_{it-1}:

$$Y_{it-1} = u_c + \phi_{sgt} + w_{it} \quad (7)$$

The u_c are teacher-grade-year effects in school s and ϕ_{sgt} are school-grade-year effects. Schools in which the null hypothesis of no systematic differences across classrooms is rejected are presumed to exhibit sorting on dimensions associated with their lagged value of Y_{it} ($H_0: u_c = 0\ \forall c$).

Horvath (2015) found more than half of all schools in North Carolina exhibited sorting on prior achievement. Additionally, she found a somewhat smaller share appeared to actively balance the gender and race composition of classrooms, and a larger share exhibited within-school sorting by parental education. (The latter is consistent with college-educated parents having an influence on their child's classroom assignment).

Schools may exhibit non-random sorting of students within grades and years, but not persistently "match" groups of students to specific teachers over time. To test for teacher matching, Horvath (2015) regresses mean lagged outcomes on ϕ_{sgt} and teacher, rather than classroom, dummies. Schools in which the null hypothesis of no teacher matching (the teacher dummies are jointly zero) is rejected are presumed to persistently match students to teachers. She finds roughly 40% of North Carolina schools persistently assigned students to teachers in this manner.

We replicated these tests for non-random classroom and teacher assignment in our NYC data, using lagged achievement and height as potential sorting variables. This test sheds light on whether classrooms within schools are grouped on height—or unobserved factors related to height—that might explain teacher effects on this outcome.

4.2 Permutation tests

Finally, we performed a series of permutation tests to provide a benchmark for what teacher "effects" look like in our data under random assignment of student data to teachers, holding class sizes constant. For these tests, we randomly allocated students to teachers in our data set within grade and year and re-estimated each baseline model. This random permutation was repeated 499 times, maintaining the actual number of students assigned to each teacher in each permutation. On each iteration, we saved the estimated standard deviation of teacher effects ($\widehat{\sigma_u}$) and then examined the distribution of these estimates across all 499 iterations.

These results were used as a Fisher exact permutation test to assess whether our estimates of the dispersion in teacher effects in the observed data differ from what one would expect under the null of no effects, with any variation necessarily caused by different classroom sizes within schools and sampling variation. Through randomly assigning students to teachers in our data, we effectively impose the null hypothesis of no sorting, no true teacher effect, no peer effects, and no systematic measurement error. If the estimated standard deviation from the observed data is larger than the 95th percentile of standard deviations from the permutations, we conclude that the standard deviation is statistically different from the null of zero.

For comparison purposes, we also report the results from permutation tests within schools. In this case we randomly allocated students to teachers within the same school and year. This imposes the null hypothesis of no sorting within schools, but between-school sorting remains possible. In this approach, there is a greater possibility that students are randomly allocated to their actual teacher, especially in smaller schools.

5 Results

5.1 Teacher effects on achievement and height

Our baseline estimates of teacher effects on the achievement and height of 4th and 5th graders in NYC are summarized in Panel A of Table 4 and visually in Figures 3-4. Panel A reports estimated standard deviations of teacher effects on each outcome. Separate standard deviations are reported for each outcome, grade, and model assumption (random or fixed effects, and with or without school effects). The random effects here are the best linear unbiased predictors (BLUPs); fixed effect coefficients were analogously adjusted by applying the shrinkage factor shown in equation 3.

Focusing first on the models without school effects, Table 4 shows that the standard deviation of teacher effects on height is substantial in NYC, despite the implausibility of a true causal effect on height. For instance, a 1σ increase in teachers' "value-added" on height is associated with a $0.21\sigma - 0.22\sigma$ increase in height in the random effects model. The adjusted fixed effects yield somewhat larger values, from 0.25σ in 4th grade to 0.32σ in 5th grade. In every case, a test of the null hypothesis that the teacher effects are jointly zero is soundly rejected at any conventional level. To put these effects in perspective, a 0.22σ increase in height amounts to a 0.68-inch gain in stature for 4th graders and an 0.72-inch gain for 5th graders. This is roughly a third of a standard deviation in year-to-year growth for children of this age.

The standard deviations of teacher effects on height are smaller, but not substantially different in standard deviation units, from those estimated for mathematics and ELA achievement. In math we find a 1σ increase in teacher value-added to be associated with a $0.29\sigma - 0.34\sigma$ and $0.25\sigma - 0.26\sigma$ increase in achievement for 4th and 5th graders, respectively. In ELA, the variation in effects is closer to $0.26\sigma - 0.28\sigma$ in 4th grade and $0.21\sigma - 0.24\sigma$ in 5th grade. All estimates of the standard deviation in teacher effects are on the upper end of the range of those found in other studies, but close to those estimated by others using NYC data (e.g., Rockoff et al., 2012).

Figures 3-4 provide full pictures of the distribution of teacher effects on height and mathematics (see supplemental appendix Figure A.3 for ELA). Both distributions are approximately symmetric around

zero, and there is generally less dispersion visible in the effects on height than in the effects on math. Comparing the 10th, 25th, 75th, and 90th percentiles of random effects in 4th grade, the centiles of teacher effects on height tend to be closer to zero (-0.22, -0.10, +0.10, and +0.21) than the same centiles for math (-0.31, -0.19, +0.15, and +0.34). The distribution of height effects is somewhat left-skewed, and the distribution of math effects somewhat right-skewed. There are a handful of relatively extreme values ($> 1.5\sigma$) in the distribution of height effects—more so than in the distribution of math effects—but fewer than 10 in total (out of 4,262 teachers). Also recall that students with outlier values for height were omitted from the analytic sample. The standard deviation of teacher effects on height, therefore, does not appear to be inflated by a small number of outliers.

The next two rows of Table 4 report similar estimates of dispersion in teacher effects for models with school effects. The inclusion of school effects should account for systematic differences in height across teachers due to school-level factors, such as differences in Fitnessgram timing, procedures for carrying out height measurements, and the like. In these cases, the estimated standard deviations are approximately 70-75 percent of those estimated in models without school effects. In all cases, however, the apparent effect of teachers on height remains meaningful in size, ranging from 0.16σ to 0.17σ. Teacher effects on math and ELA are comparably reduced when including school effects (to $0.15\sigma - 0.22\sigma$).

5.2 Do teacher effects on height imply bias for teacher effects on achievement?

A possible interpretation for non-trivial "effects" of teachers on height is that students are sorted to teachers on unobserved factors related to height. These factors might include student health, ethnicity or immigration history, past grade retention patterns, red-shirting, or birthweight, for example. If these unobserved factors are also related to achievement, this would be a potentially troubling finding for achievement VAMs. That is, if commonly used covariates in value-added models for achievement insufficiently capture this sorting. In their review of the literature; Koedel, Mihaly, & Rockoff (2015) find that value-added measures for achievement are not especially sensitive to the choice of covariates. While

plausible, the sorting mechanism would have to be more complicated here, since our height models condition on lagged height. Sorting would effectively have to occur on grade-to-grade changes in height rather than levels.

To explore this possibility, we first examined how teachers' estimated effects on achievement correlate with their effects on height. These correlations are reported in Table 5 for the baseline models. There is little evidence of an association between teacher effects on height and academic achievement. In both 4th and 5th grade, we find the correlation between value-added on height and achievement is typically smaller than 0.05 in absolute value. We found similar results when using Spearman rank correlations. In contrast, the correlation between teachers' value-added on math and ELA achievement is modest to strong, at 0.64-0.70 in 4th grade, and 0.51-0.56 in 5th grade. There were only two significant correlations between effects on height and achievement: the positive correlations of 0.199 and 0.090 between height and math effects in 4th and 5th grades, respectively. Both emerge only when teacher effects are estimated as fixed effects; the corresponding correlations for random effects are close to zero. We have examined these two cases closely and have been unable to identify alternative explanations for these associations. For example, there are no outlier fixed effect estimates that drive up the correlation. When we exclude effects with an absolute value of 1.5 or higher, the correlation remains unchanged. The correlation is also not attributable to the shrinkage factor applied to the fixed effect estimates; the correlation is present before the adjustment. While not shown in this table, we also found positive correlations between value-added on height and ELA, but smaller, and again only with the fixed effects model.

The low correlations in Table 5 offer some assurance that the estimated effects on height are not evidence of sorting on factors related to achievement. They do not, however, rule out the possibility that students sort to classrooms or teachers on factors related to stature—or changes in stature—that are *unrelated* to achievement. To examine this, we began by using Horvath's (2015) method to identify schools that exhibit non-random sorting of students to classrooms on prior characteristics, including height. Such grouping might be indicative of sorting on unobserved factors related to height. As described

in Section 4.3, this involved estimating separate regressions for each school to test the null hypothesis of no mean differences in students across classrooms. As in that study, we obtained *p*-values for each school, separately by grade, and separately for height and math (ELA results are similar). *p*-values below 0.05 are interpreted as evidence of systematic sorting across classrooms within school-years.

Results from these tests are shown in Figure 5. The histograms in this figure show the relative frequency of *p*-values across schools, separately by grade level and outcome. We find strong evidence of classroom grouping based on lagged math achievement, but little evidence of grouping on height. For the roughly 700 schools in the math regressions, we can reject the null hypothesis of no sorting in 64.6 percent of cases in 4th grade, and 62.6 percent in 5th grade. These proportions are remarkably close to those reported in Horvath (2015), who estimated that 60 percent of North Carolina schools exhibited systematic sorting on prior achievement. In contrast, of the 680 schools in the height regressions, we can reject the null hypothesis in only 10.1 percent of cases in 4th grade, and 11.2 percent in 5th grade. This is more than would be predicted by chance, but a much lower prevalence of rejections relative to math.

Evidence of classroom grouping was more consistent across grades in math than in height, a pattern one would expect if grouping were a school-level practice for achievement but not height. For example, conditional on rejecting the null hypothesis of no sorting in 4th grade, a school had an 85.7 percent chance of rejection in 5th grade (as compared to 17.2 percent for those that did not appear to sort in 4th grade). For height, the comparable numbers were 24.6 and 9.9 percent. Only 2.5 percent of schools appeared to sort students on height in both grades.

We also conducted Horvath's test for teacher matching, that is, persistent sorting of students to teachers across years. In this case we found 32.9 percent of schools appeared to match students to teachers based on math scores (compared to 40% in Horvath's study), while only 8.1 percent appeared to match based on height. These results are available in supplemental appendix Figure A.4.

Our second approach to testing for systematic sorting on unobserved characteristics associated with height was to estimate teacher-by year effects u_{jt} and examine how these effects correlate over time for the same teacher. If there is a persistent “effect” of teachers on height, potentially explained by teacher

matching, one would expect to see a positive correlation in classroom effects for the same teacher over time. Instead, we find this correlation is small in absolute value and close to zero, as reported in Table 6. The between-year correlations in teacher effects on height are negative (about -0.166) in the random effects model, but in the fixed effects model range from 0.001 in 4th grade to -0.094 in 5th grade. By contrast, the intertemporal correlations are 0.435-0.587 in math, depending on the model assumptions and grade, and 0.210-0.501 in ELA.

5.3 Are teacher effects on height entirely "noise"?

The analysis thus far finds little evidence in support of systemic sorting of students to teachers on unobserved factors related to height. While a majority of NYC schools exhibit non-random sorting of students to classrooms on prior achievement, only a small proportion exhibit such sorting on height (or unobserved factors related to height). Moreover, teacher effects from our baseline models for height show little to no persistence across years, when correlating effects for teachers with multiple years of classroom data. This finding would not preclude classroom-level sorting, but it is not consistent with persistent matching of students to teachers across years.

As described in Section 4.2, an alternative to the baseline model explicitly allows for a random classroom effect that is uncorrelated within teachers over time. This is the approach used by Kane, Staiger, and Rockoff (2008) in estimating teacher effects as mean classroom-level residuals multiplied by a shrinkage factor. (They can also be estimated using maximum likelihood and a 3-level random effects model.) The critical difference in this approach is that the “signal” component of the shrinkage factor is estimated from the covariance in classroom effects for the same teacher in successive years. Our results in the previous section suggested this covariance is close to zero for these estimates of teacher effects on height. The shrinkage factor using this method would be the theoretically “correct” one, if there were no persistent teacher effects on height.

Panels (B) and (C) of Table 4 report the estimated standard deviations in teacher effects on achievement and height when fitting 3-level models. The estimates in panel (B) come from the mean

residuals approach, while those in panel (C) come from maximum likelihood. In these cases, the standard deviation in teacher effects on height falls to zero, while those for math and ELA remain significant, ranging from 0.087-0.199, depending on the model, grade, and subject. In cases where the covariance in annual teacher effects was negative, we set σ_u to zero. These results suggest that the estimated effects in our baseline models for height are almost certainly not "true" effects on height.

To assess the likelihood that differences in sample sizes of classrooms within schools plus sampling variation could produce teacher effects like those observed in the baseline model, we removed all effects of sorting, peers, systematic measurement error, and true effects, and randomly assigned students to teachers as described in Section 4.4. This random assignment was repeated 499 times for each model, and the estimated standard deviation ($\widehat{\sigma_u}$) was retained on each iteration. Random effect estimation using maximum likelihood did not converge when student data was randomly assigned to teachers. Thus, for the random effect models, we calculated mean residuals for teachers and multiplied by the shrinkage factor. Distributions of the $\widehat{\sigma_u}$ across permutations are shown in Figure 6, and the means of these distributions are reported in Panel D of Table 4.

Even under random permutations of student data to teachers, we continue to find meaningful (in magnitude) effects of teachers on height and achievement in the baseline models although we can reject the null that the true SD is zero. In the fixed effects models, the average $\widehat{\sigma_u}$ across permutations ranged from 0.053 for height to 0.068 for ELA. The standard deviation of these estimates across permutations is roughly 0.001-0.002. In other words, even when (real) data on students is randomly allocated across teachers, a 1σ increase in teacher "value-added" is associated with a 0.053σ increase in height and a 0.068σ increase in ELA test performance. As one would expect, teacher effects under random assignment of students to teachers are uncorrelated with those estimated with the actual data. (They are also uncorrelated across subjects.) Moreover, permutation tests that use the 3-level model produce a $\widehat{\sigma_u}$ close to zero.

We repeated this permutations test by allocating students to teachers at random within schools. Distributions of the $\widehat{\sigma_u}$ are shown in Figure 6, and the means are reported in Panel E of Table 4. The average $\widehat{\sigma_u}$ across permutations within school is larger (0.072-0.131) in this case, which is not surprising since this method may not eliminate systematic measurement error between schools (and some students will be randomly matched to their actual teacher when randomizing within school).

The permutation test offers two important insights. First, even under completely random assignment of student data to teachers, there are nonzero teacher "effects" in our baseline model. Even more important, the distribution of effects under random permutations provides a sense of the range of standard deviations under an imposed null of no systematic sorting, peer effects, or true effects. Second, our estimated teacher effects in the observed data are clearly over-dispersed relative to this null effect distribution, suggesting our estimated teacher effects on height and achievement have a standard deviation that is statistically different from 0. To take one example, the 95th percentile of the $\widehat{\sigma_u}$ for 4th grade height among the permutations is 0.06 (Figure 6). This can be compared to an estimated $\widehat{\sigma_u}$ in the actual data of 0.218. Similar differences are observed in math and ELA. In the case of height, this suggests that there is some systematic variation beyond randomness associated with the actual class sizes within schools and sampling variation in the covariates.

6 Discussion

Schools and districts across the country want to employ teachers who can best help students to learn, grow, and achieve academic success. Identifying such individuals is integral to schools' success but is also difficult to do in practice. In the face of data and measurement limitations, school leaders and state education departments seek low-cost, unbiased ways to observe and monitor the impact that their teachers have on students. Although many have criticized the use of VAMs to evaluate teachers, they remain a widely-used measure of teacher performance. In part, their popularity is due to convenience-while observational protocols which send observers to every teacher's classroom require expensive training and considerable resources to implement at scale, VAMs use existing data and can be calculated centrally at

low cost. Further, VAMs are arguably less biased than many other evaluation methods that districts might use instead (Bacher-Hicks et al. 2017; Harris et al. 2014; Hill et al. 2011).

Yet questions remain about the reliability, validity, and practical use of VAMs. This paper interrogates concerns raised by prior research on VAMs and raises new concerns about the use of VAMs in career and compensation decisions. We explore the bias and reliability of commonly estimated VAMs by comparing estimates of teacher value-added in mathematics and ELA with parallel estimates of teacher value-added on a well-measured biomarker that teachers should not impact: student height. Using administrative data from New York City, we find estimated teacher "effects" on height that are comparable in magnitude to actual teacher effects on math and ELA achievement, 0.22σ compared to 0.29σ and 0.26σ respectively. On its face, such results raise concerns about the validity of these models.

Fortunately, subsequent analysis finds that teacher effects on height are primarily noise, rather than bias due to sorting on unobserved factors. To ameliorate the effect of sampling error on value-added estimates, analysts sometimes "shrink" VAMs, scaling them by their estimated signal-to-noise ratio. When we apply the shrinkage method across multiple years of data from Kane and Staiger (2008), the persistent teacher "effect" on height goes away, becoming the expected (and known) mean of zero. This procedure is not always done in practice, however, and requires multiple years of classroom data for the same teachers to implement. Of course, for making hiring and firing decisions, it seems important to consider that value added measures which require multiple years of data to implement will likely permit identification of persistently bad teachers, but not provide a performance evaluation metric that can be met by teachers trying to improve their performance. In more realistic settings where the persistent effect is not zero, it is less clear that shrinkage would have a major influence on performance decisions, since it has modest effects on the relative rankings of teachers.

Taken together, our results provide a cautionary tale for the naïve application of VAMs to teacher evaluation and other settings. They point to the possibility of the misidentification of sizable teacher "effects" where none exist. These effects may be due in part to spurious variation driven by the typically small samples of children used to estimate a teacher's individual effect.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools," Journal of Labor Economics 25: 95-135.

American Institutes for Research. 2013. 2012-2013 Growth Model for Educator Evaluation: Technical Report Prepared for the New York State Education Department. Washington, D.C.: American Institutes for Research.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. 2017. "An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys." National Bureau of Economic Research Working Paper Series, No. 23478.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. 2014. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." National Bureau of Economic Research Working Paper Series, No. 20657.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. 2010. Problems with the Use of Student Test Scores to Evaluate Teachers. Economic Policy Institute Briefing Paper No. 278.

Ballou, D., Mokher, C. G., and Cavalluzzo, L. 2012. "Using Value-Added Assessment for Personnel Decisions: How Omitted Variables and Model Specification Influence Teachers' Outcomes." Working paper.

Ballou, D., Sanders, W., and Wright, P. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." Journal of Educational and Behavioral Statistics, 29: 37-65.

Braun, H. I., Chudowsky, N., & Koenig, J. 2010. Getting Value Out of Value-Added: Report of a Workshop. Washington, D.C.: National Academies Press.

Buddin, Richard, and Gema Zamarro. 2009. "Teacher Qualifications and Student Achievement in Urban Elementary Schools," Journal of Urban Economics 66: 103-115.

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

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

Doherty, Kathryn M., and Sandi Jacobs. 2013. Connect the Dots: Using Evaluations of Teacher Effectiveness to Inform Policy and Practice. Washington, D.C.: National Council on Teacher Quality.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., and Whitehurst, G. 2010. “Evaluating Teachers: The Important Role of Value-Added.” Washington, D.C.: Brookings Institution.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., and Whitehurst, G. J. 2011. “Passing Muster: Evaluating Teacher Evaluation Systems.” Washington, D.C.: Brookings Institution.

Goldhaber, D., & Chaplin, D. 2011. “Assessing the 'Rothstein Test': Does it Really Show Teacher Value-Added Models are Biased?” Washington, D.C.: CALDER Center Working Paper No. 71.

Goldhaber, D., and Hansen, M.L. 2012. “Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance,” Economica 80: 589-612.

Goldhaber, D. D., Goldschmidt, P., and Tseng, F. 2013. “Teacher Value-Added at the High-School Level: Different Models, Different Answers?” Educational Evaluation and Policy Analysis, 35: 220-236.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. “Identifying Effective Teachers Using Performance on the Job.” Washington, D.C.: Brookings Institution.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P., and Wooldridge, J. M. 2015. “An Evaluation of Empirical Bayes’ Estimation of Value-Added Teacher Performance Measures.” Journal of Educational and Behavioral Statistics, 40(2), 190-222.

Hanushek, Eric A. 2009. “Teacher Deselection,” in Creating a New Teaching Profession, Dan Goldhaber and Jane Hannaway (eds.) Washington, D.C.: The Urban Institute Press.

Hanushek, E. A., and Rivkin, S. G. 2010. “Generalizations about Using Value-Added Measures of Teacher Quality.” American Economic Review, 100: 267-271.

Harris, Douglas, and Tim R. Sass. 2006. “Value-Added Models and the Measurement of Teacher Quality,” Working Paper, Florida State University.

Harris, Douglas N. 2011. Value-Added Measures in Education What Every Educator Needs to Know. Cambridge: Harvard Education Press.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. 2014. “How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures.” American Educational Research Journal, 51: 73-112.

Hill, H. C., Kapitula, L.,& Umland, K. 2011. A Validity Argument Approach to Evaluating Teacher Value-Added Scores. American Educational Research Journal, 48: 794-831.

Horvath, Hedvig. 2015. “Classroom Assignment Policies and Implications for Teacher Value-Added Estimation.” Job Market Paper, University of California-Berkeley Department of Economics.

Isenberg, E., and Hock, H. 2011. Design of Value-Added Models for IMPACT and TEAM in DC Public Schools, 2010–2011 School Year: Final Report. Washington, DC: American Institutes for Research.

Ishii, J., and Rivkin, S. G. 2009. “Impediments to the Estimation of Teacher Value Added.” Education Finance and Policy, 4: 520-536.

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. 2013. “Teacher Effects and Teacher-Related Policies.” Annual Review of Economics.

Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. 2013. “Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.” Seattle, WA: Gates Foundation.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. “What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City,” Economics of Education Review 27: 615-631.

Kane, Thomas J., and Douglas O. Staiger. 2008. “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation,” National Bureau of Economic Research Working Paper #14607.

Kinsler, J. 2012. “Assessing Rothstein's Critique of Teacher Value-Added Models. Quantitative Economics, 3(2), 333-362.

Koedel, C., and 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, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. “The Intertemporal Variability of Teacher Effect Estimates,” Education Finance and Policy 4: 572-606.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” American Economic Review, Papers and Proceedings of the American Economics Association. 94: 247-252.

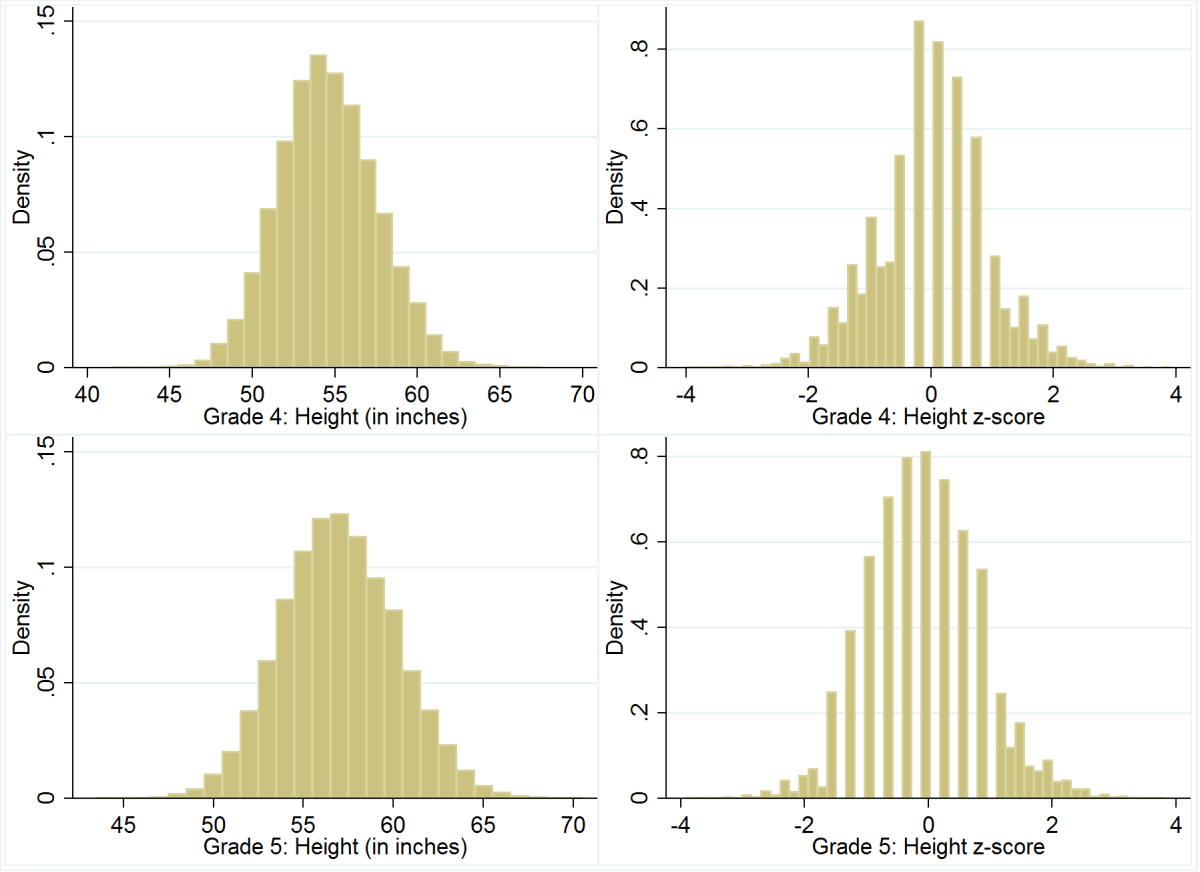
Rockoff, J. E., Staiger, D. O., Kane, T. J., and Taylor, E. S. 2012. “Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools.” American Economic Review, 102(7), 3184–-3213.

Rothstein, Jesse. 2010. “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.” Quarterly Journal of Economics 125: 175-214.

Schochet, Peter Z. and Hanley S. Chiang. 2013. “What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?” Journal of Educational and Behavioral Statistics, 38: 142-171.

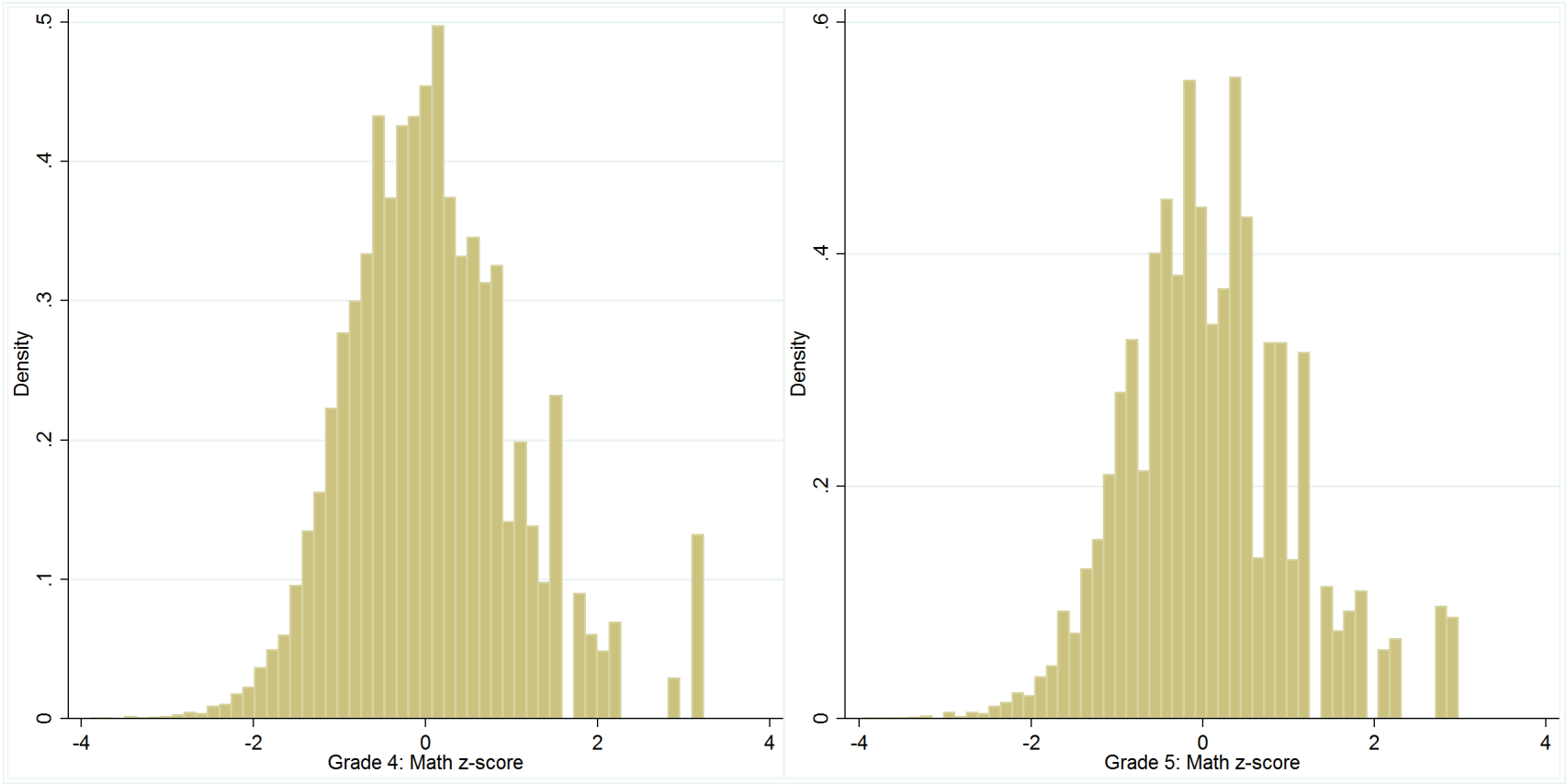
Weisburg, Daniel, Susan Sexton, Susan Mulhern, and David Keeling. 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. The New Teacher Project.

Figure 1: Distributions of height for NYC 4th and 5th graders



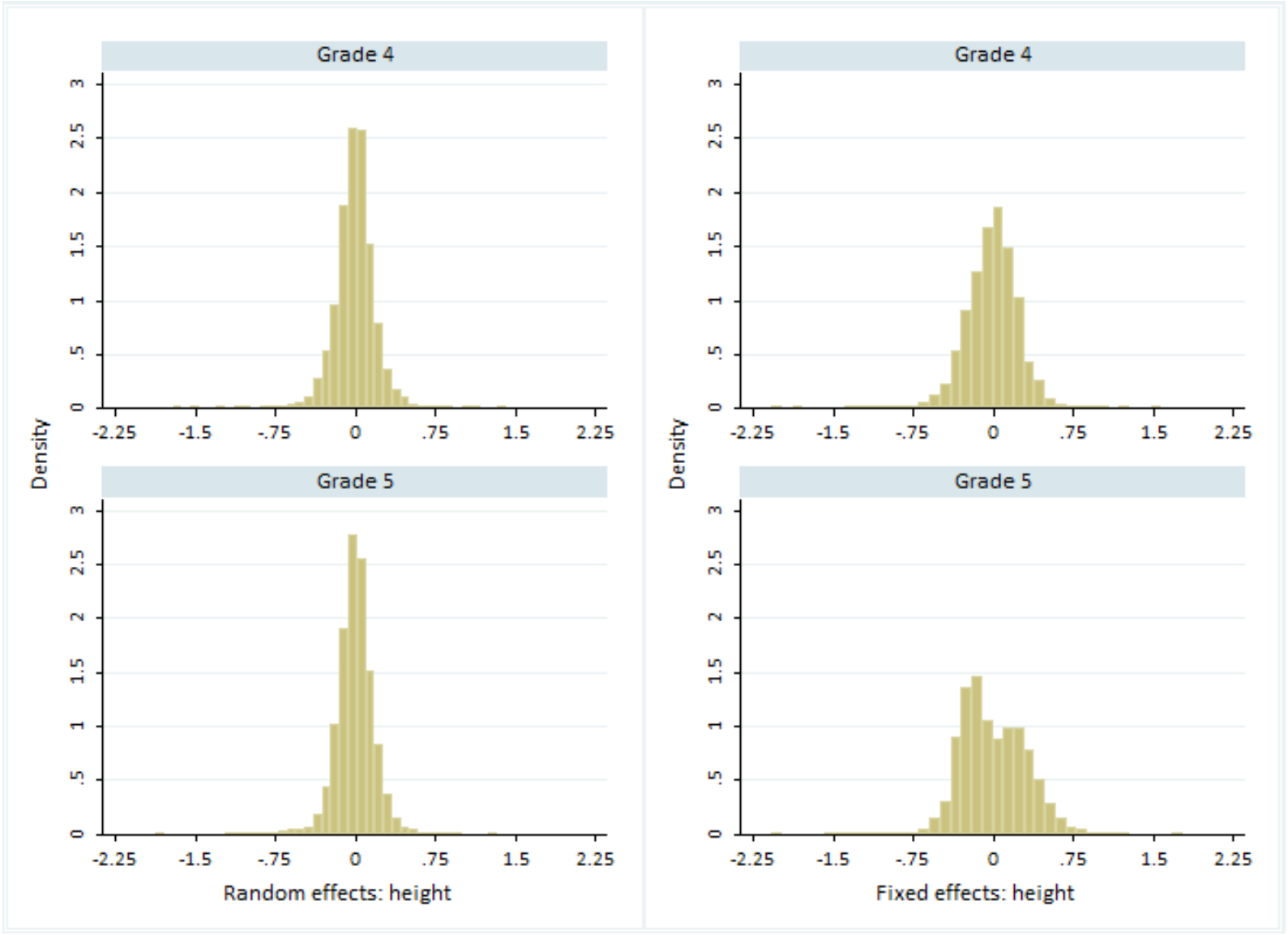
Notes: height in inches takes on 26 unique values in 4th grade and 27 unique values in 5th grade. The z-score for height (standardized by grade and year) takes on 98 unique values in 4th grade and 103 unique values in 5th grade (but only 26-27 unique values in each year).

Figure 2: Distributions of math scores for NYC 4th and 5th graders



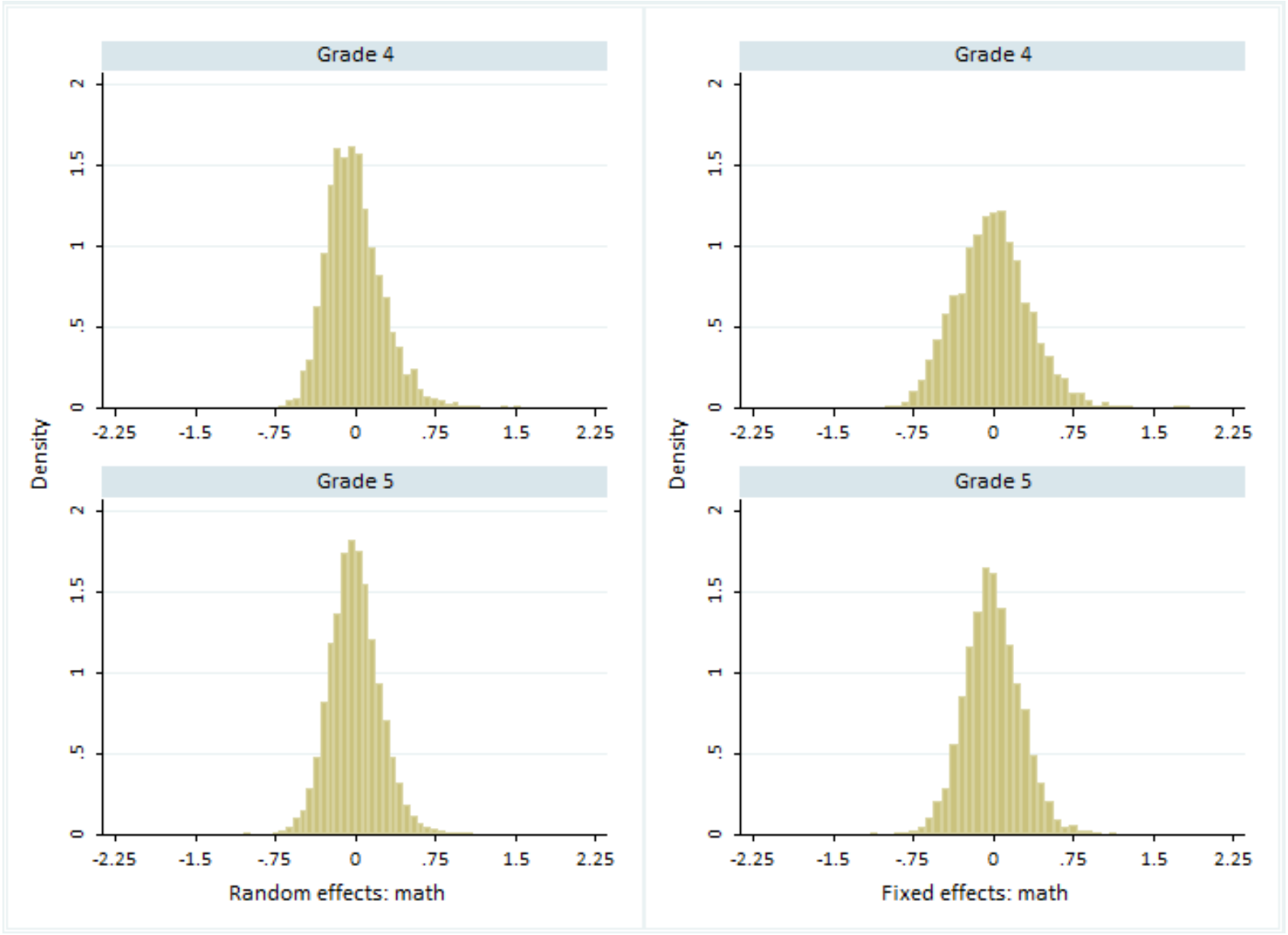
Notes: over all years, the z-score for math (standardized by grade and year) takes on 261 unique values in 4th grade and 169 unique values in 5th grade. The number of unique values in each year is approximately 65 in 4th grade and 43 in 5th grade. (There are more unique values overall, as the mean and standard deviation of scale scores varies by year).

Figure 3: Distribution of teacher effects on height



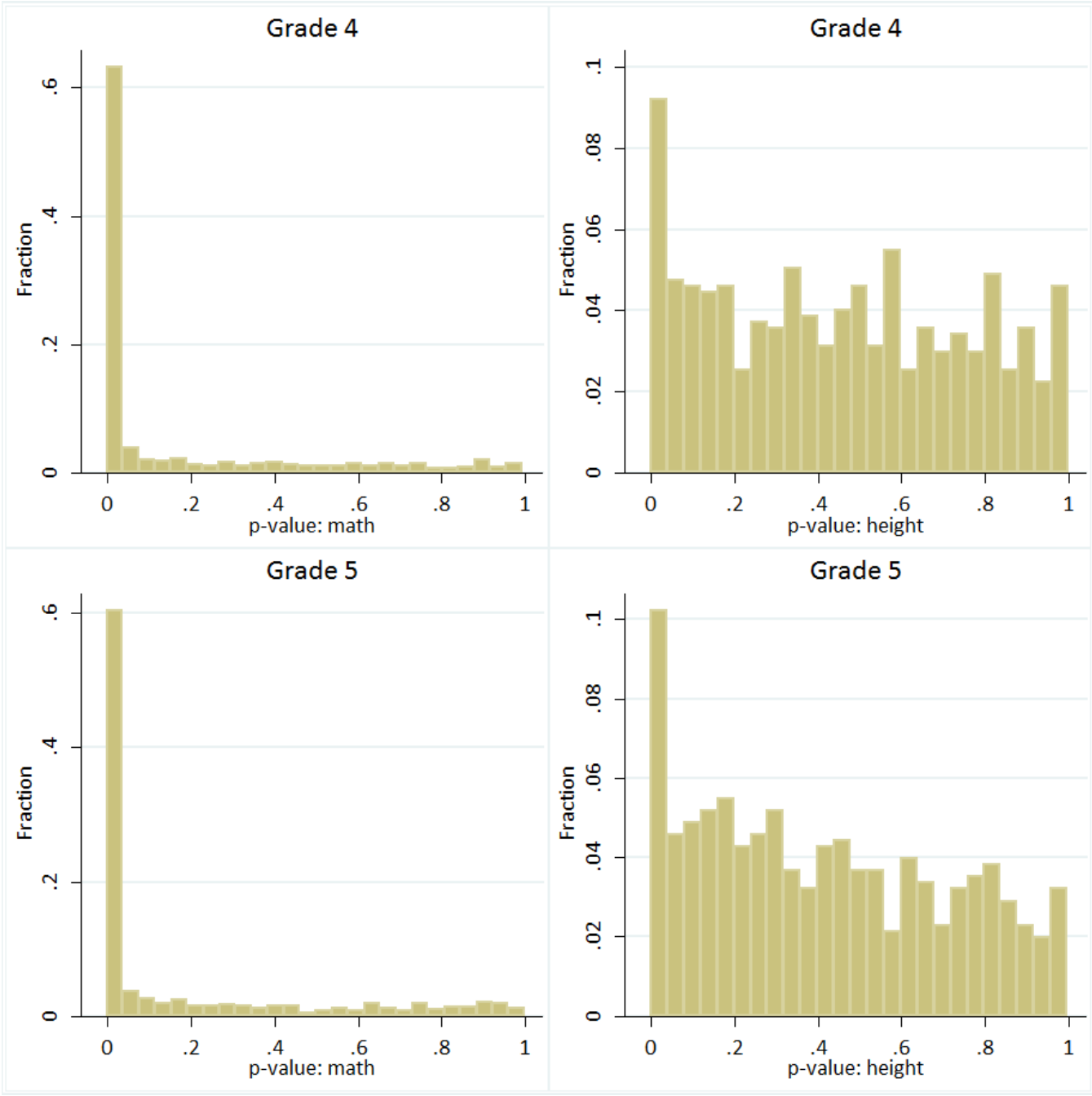
Notes: see notes to Table 4 for a description of how teacher effects were estimated. N=4,262 4th grade teachers and 3,687 5th grade teachers.

Figure 4: Distribution of teacher effects on math achievement



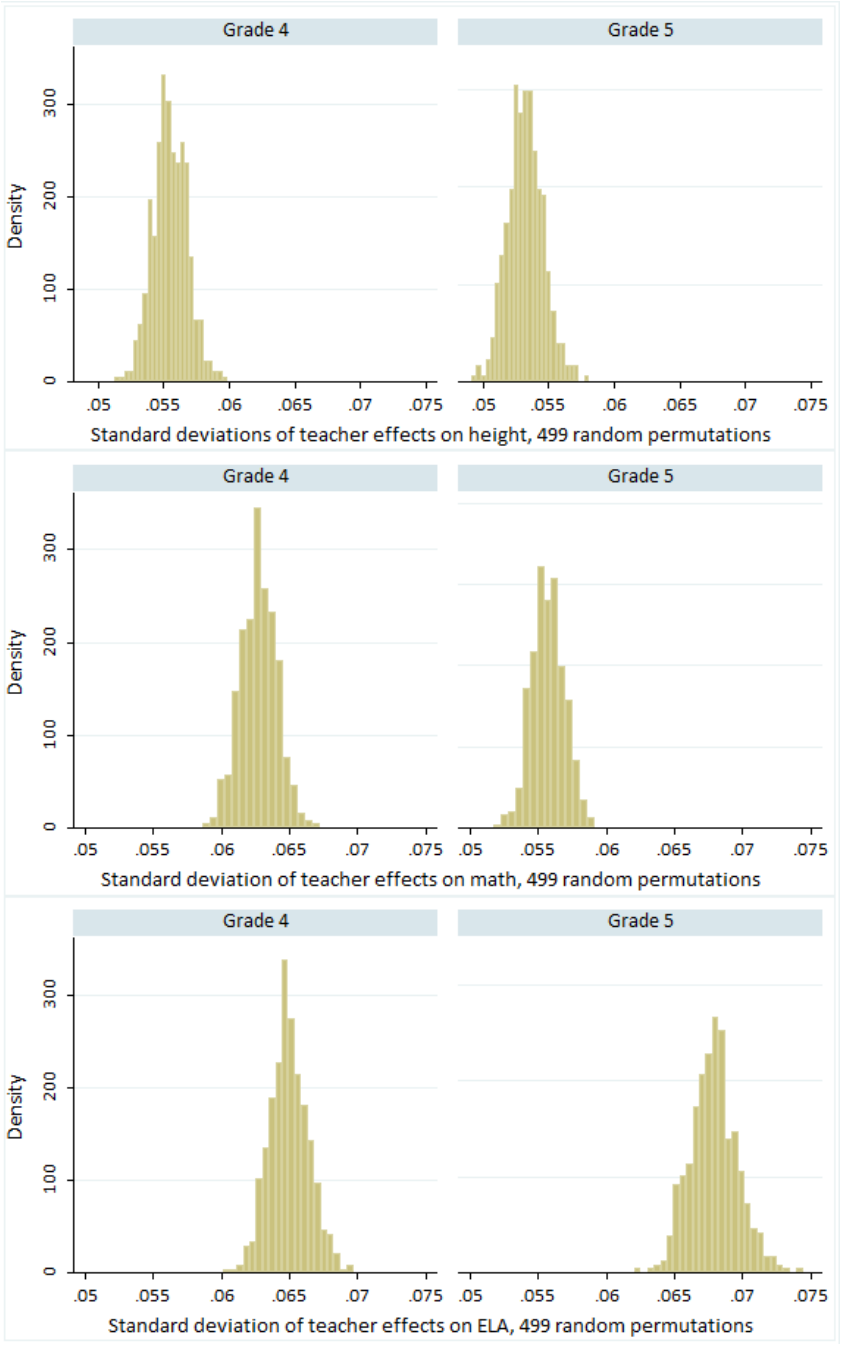
Notes: see notes to Table 4 for a description of how teacher effects were estimated. N=4,721 4th grade teachers and 4,249 5th grade teachers.

Figure 5: Tests for nonrandom tracking by prior math achievement and height



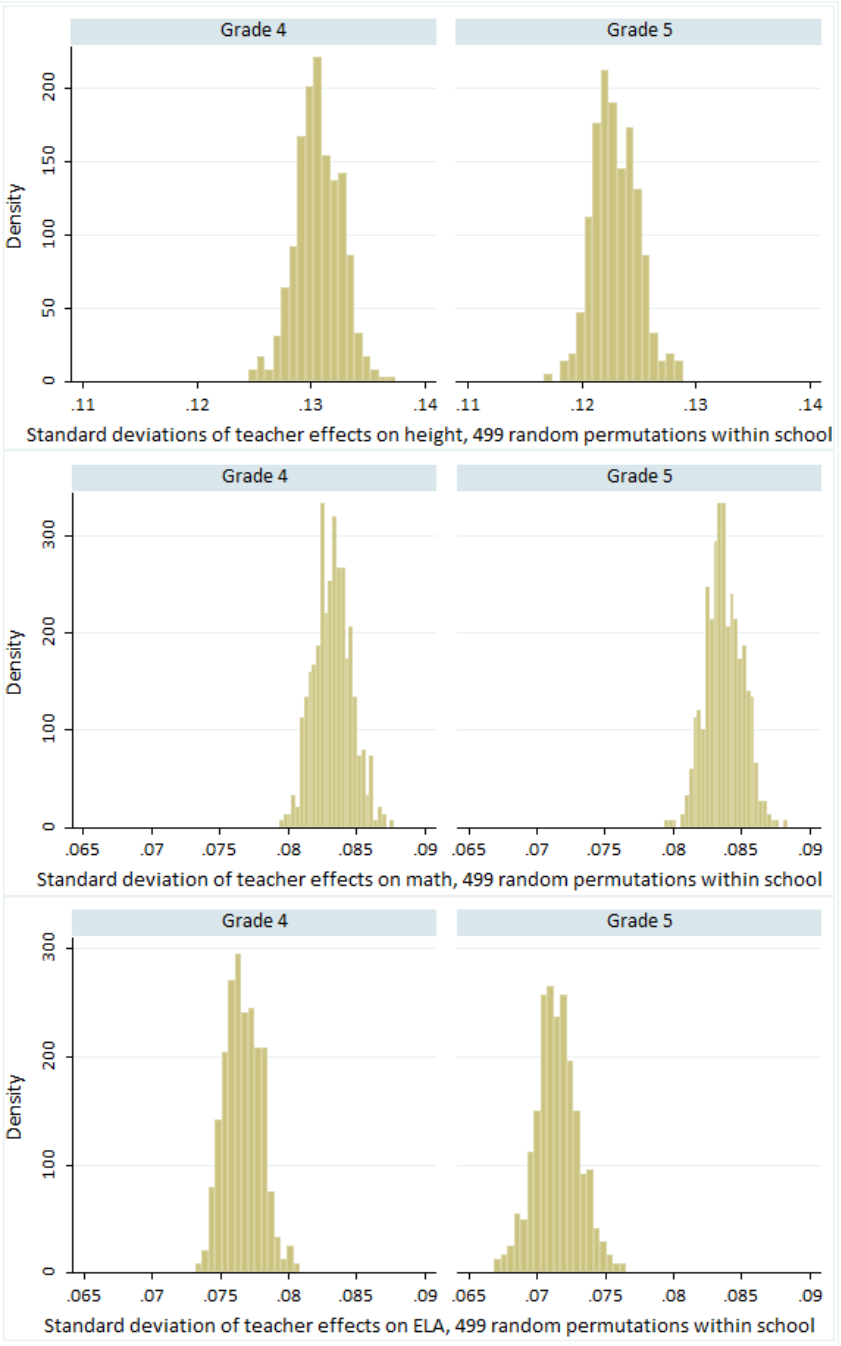
Notes: Each p-value is from a test of the hypothesis that classroom effects in a school s are jointly zero. Regression models are estimated separately for each school and grade, with lagged student outcomes regressed on school-grade-year and classroom effects.

Figure 6: Standard deviations of teacher effects from 499 random permutations



Notes: to create these figures, we repeated the following steps 499 times. First, randomly allocate all students in our data to teachers (within year). Then, re-estimate the value-added model assuming fixed or random effects. (Standard deviations of the adjusted fixed effects are shown). For each iteration we saved the estimated standard deviation in teacher effects $\widehat{\sigma_u}$. These figures show the distribution of $\widehat{\sigma_u}$'s across permutations.

Figure 7: Standard deviations of teacher effects from 499 random permutations within schools



Notes: to create these figures, we repeated the following steps 499 times. First, randomly allocate all students in our data to teachers (within *school* and year). Then, re-estimate the value-added model assuming fixed or random effects. (Standard deviations of the adjusted fixed effects are shown). For each iteration we saved the estimated standard deviation in teacher effects $\widehat{\sigma_u}$. These figures show the distribution of $\widehat{\sigma_u}$'s across permutations.

Table 1: Mean student characteristics, analytic samples and all students

	Grade 4			Grade 5		
	All linked obs	Height sample	Math sample	All linked obs	Height sample	Math sample
ELA z-score	0.027	0.072	0.051	0.025	0.069	0.047
Math z-score	0.033	0.103	0.079	0.035	0.119	0.087
Height z-score	-0.008	-0.034	-0.012	-0.007	-0.031	-0.007
Height (inches)	54.662	54.585	54.649	57.082	57.001	57.080
Female	0.506	0.509	0.507	0.505	0.507	0.507
White	0.156	0.169	0.162	0.152	0.167	0.157
Black	0.283	0.275	0.281	0.286	0.277	0.283
Hispanic	0.392	0.376	0.386	0.395	0.376	0.388
Asian	0.162	0.181	0.171	0.162	0.181	0.171
Age	9.645	9.626	9.640	10.670	10.646	10.665
Low income	0.798	0.804	0.804	0.799	0.805	0.806
LEP	0.119	0.102	0.105	0.101	0.082	0.086
Special ed	0.119	0.115	0.118	0.116	0.111	0.114
English at home	0.585	0.576	0.582	0.573	0.564	0.570
Recent immigrant	0.130	0.117	0.117	0.148	0.137	0.137
Same math/ELA teacher	0.900	0.883	0.893	0.862	0.858	0.867
Manhattan	0.133	0.119	0.125	0.131	0.115	0.125
Bronx	0.207	0.165	0.184	0.209	0.158	0.181
Brooklyn	0.312	0.340	0.325	0.310	0.348	0.328
Queens	0.284	0.302	0.299	0.287	0.304	0.299
Staten Island	0.064	0.074	0.068	0.063	0.074	0.068
2007	0.241	0.167	0.204	0.247	0.174	0.207
2008	0.243	0.225	0.241	0.244	0.235	0.242
2009	0.251	0.274	0.259	0.254	0.279	0.261
2010	0.264	0.334	0.297	0.255	0.311	0.290
N	239,577	153,242	182,623	236,983	143,738	180,639

Notes: "All linked observations" refers to all students who could be linked to their classroom teacher. The analytic samples include all students who meet the minimum data requirements to be included in the teacher value-added models for height or math. "Same math/ELA teacher" means the student has the same teacher code reported for math and ELA. In some cases the teacher code differs because a teacher code is not reported for either math or ELA (more common for ELA, since not all students take this test). When conditioning on non-missing math and ELA teacher codes, the percent of 4th graders with the same math and ELA teacher code exceeds 98%; the percent of 5th graders exceeds 96%.

Table 2: Count of unique teachers and classrooms, and students per teacher or classroom in analytic sample

	Height		Math		ELA	
	Grade 4	Grade 5	Grade 4	Grade 5	Grade 4	Grade 5
Unique teachers (N)	4,263	3,687	4,721	4,249	4,366	3,978
Mean years observed	1.90	1.98	1.88	1.94	1.82	1.87
Students per teacher:						
Mean	36.0	39.0	38.7	42.5	35.9	39.5
SD	22.9	25.5	24.5	27.4	22.8	24.9
p25	19	20	20	21	19	20
p50	27	29	29	33	26	29
p90	71	76	77	84	72	78
Unique classrooms (N)	7,594	6,848	8,712	8,138	7,941	7,451
Students per classroom:						
Mean	20.0	20.8	20.9	22.2	19.7	21.1
SD	5.4	6.4	5.1	6.4	5.6	6.3
p25	17	17	18	19	16	18
p50	20	21	21	22	20	21
p90	26	28	27	28	26	28

Notes: Teachers and classrooms are counted only when seven or more students were available with the minimum data to be included in the value-added models for these outcomes. For the full distributions, see the supplemental appendix.

Table 3: Student-level bivariate correlations in outcome variables

Correlations between:	Grade 4	Grade 5
Math and ELA	0.688^{***}	0.585^{***}
Math and height	-0.059	-0.068^{***}
ELA and height	-0.046^{***}	-0.042^{***}
Correlation with lag:	**Grade 4**	**Grade 5**
Math	0.701^{***}	0.757^{***}
ELA	0.683^{***}	0.646^{***}
Height	0.799^{***}	0.793^{***}
Correlations between changes in:	**Grade 4**	**Grade 5**
Math and ELA	0.158^{***}	0.140^{***}
Math and height	0.002	0.007^{**}
ELA and height	0.013^{***}	-0.006^{*}

Notes: Pairwise correlations using all students with available data, not just those in the analytic VAM samples. All outcome measures are z-scores, where the height measure is standardized by grade and year. ***, **, and * indicate statistically significant correlations at the 0.0001, 0.01 and 0.05 levels, respectively.

Table 4: Standard deviation of estimated teacher effects

	Grade 4			Grade 5		
Model	Height	Math	ELA	Height	Math	ELA
A. Baseline models						
RE	0.218	0.286	0.256	0.210	0.253	0.210
FE (adj)	0.250	0.344	0.278	0.315	0.258	0.240
RE w/school effects	0.169	0.216	0.184	0.157	0.199	0.155
FE w/school effects (adj)	0.166	0.202	0.172	0.160	0.189	0.145
B. 3-level models (KS&R)						
RE	0.000	0.163	0.104	0.000	0.132	0.097
RE w/school effects	0.000	0.107	0.077	0.002	0.087	0.062
C. 3-level models (MLE)						
RE	0.000	0.199	0.159	0.000	0.164	0.121
RE w/school effects	0.000	0.108	0.070	0.000	0.089	0.056
D. Permutations						
FE (adj) - mean $\widehat{\sigma_u}$	0.056	0.063	0.065	0.053	0.056	0.068
E. Permutations within school						
FE (adj) - mean $\widehat{\sigma_u}$	0.131	0.083	0.077	0.123	0.084	0.072

Notes: For Panel A, teacher effects were estimated in four ways: (1) assuming random teacher effects; (2) assuming fixed teacher effects and "shrinking" using the estimated signal-to-noise ratio after estimation; (3) assuming random teacher effects and including school fixed effects; and (4) assuming fixed teacher effects (shrunken after estimation) and including school effects—uses a two stage method that regresses outcome on covariates and school fixed effects and then uses the residuals to estimate the teacher fixed effects. For Panel B, mean residuals for each teacher were shrunk using Equation 3, from models without and with school fixed effects. For Panel C, a 3-level random effects model was used. Panels D and E show the mean $\widehat{\sigma_u}$ across 499 permutations of students to teachers within year (Panel D) or students to teachers within schools and years (Panel E).

Table 5: Pairwise correlations between teacher effects

			Math VAM:	
Grade 4	RE	FE (adj)	RE w/ school effects	FE w/ school effects
Height VAM:				
RE	**-0.019**	-0.014	-0.007	0.008
FE (adj)	-0.030^{+}	**0.199***	-0.022	-0.023
RE w/school effects	0.000	-0.003	**0.002**	0.002
FE w/school effects	-0.002	-0.004	0.001	**0.000**
ELA VAM:				
RE	**0.697***	0.597*	0.521*	0.519*
FE (adj)	0.658*	**0.689***	0.477*	0.475*
RE w/school effects	0.525*	0.432*	**0.646***	0.643*
FE w/school effects	0.522*	0.428*	0.643*	**0.641***
Grade 5	RE	FE (adj)	RE w/ school effects	FE w/ school effects
Height VAM:				
RE	**0.016**	0.015	0.002	0.002
FE (adj)	0.009	**0.090***	0.005	0.005
RE w/school effects	0.001	0.002	**-0.006**	-0.007
FE w/school effects	0.000	0.002	0.005	**0.005**
ELA VAM:				
RE	**0.557***	0.540*	0.438*	0.434*
FE (adj)	0.511*	**0.562***	0.382*	0.378*
RE w/school effects	0.425*	0.406*	**0.514***	0.509*
FE w/school effects	0.424*	0.405*	0.514*	**0.511***

Notes: See notes to Table 4 for a description of how teacher effects were estimated. All correlations at pairwise at the teacher level. * indicates statistical significance at the 0.001 level. + indicates significance at 0.05 level.

Table 6: Between-year correlations in teacher effects

	Grade 4	Grade 5	N(4)	N(5)
Height:				
RE	-0.166	-0.167	3319	3135
FE (adj)	0.001	-0.094	3319	3135
RE w/school effects	-0.004	0.007	3285	3100
FE w/school effects (adj)	0.000	0.011	3285	3100
Math:				
RE	0.557	0.479	4001	3885
FE (adj)	0.587	0.498	4001	3885
RE w/school effects	0.463	0.435	3988	3868
FE w/school effects (adj)	0.471	0.438	3988	3868
ELA:				
RE	0.456	0.408	3428	3357
FE (adj)	0.501	0.453	3428	3357
RE w/school effects	0.247	0.210	3410	3345
FE w/school effects (adj)	0.249	0.214	3410	3345

Notes: See notes to Table 4 for a description of how teacher effects were estimated. All correlations at pairwise at the teacher level. * indicates statistical significance at the 0.001 level. + indicates significance at 0.05 level.

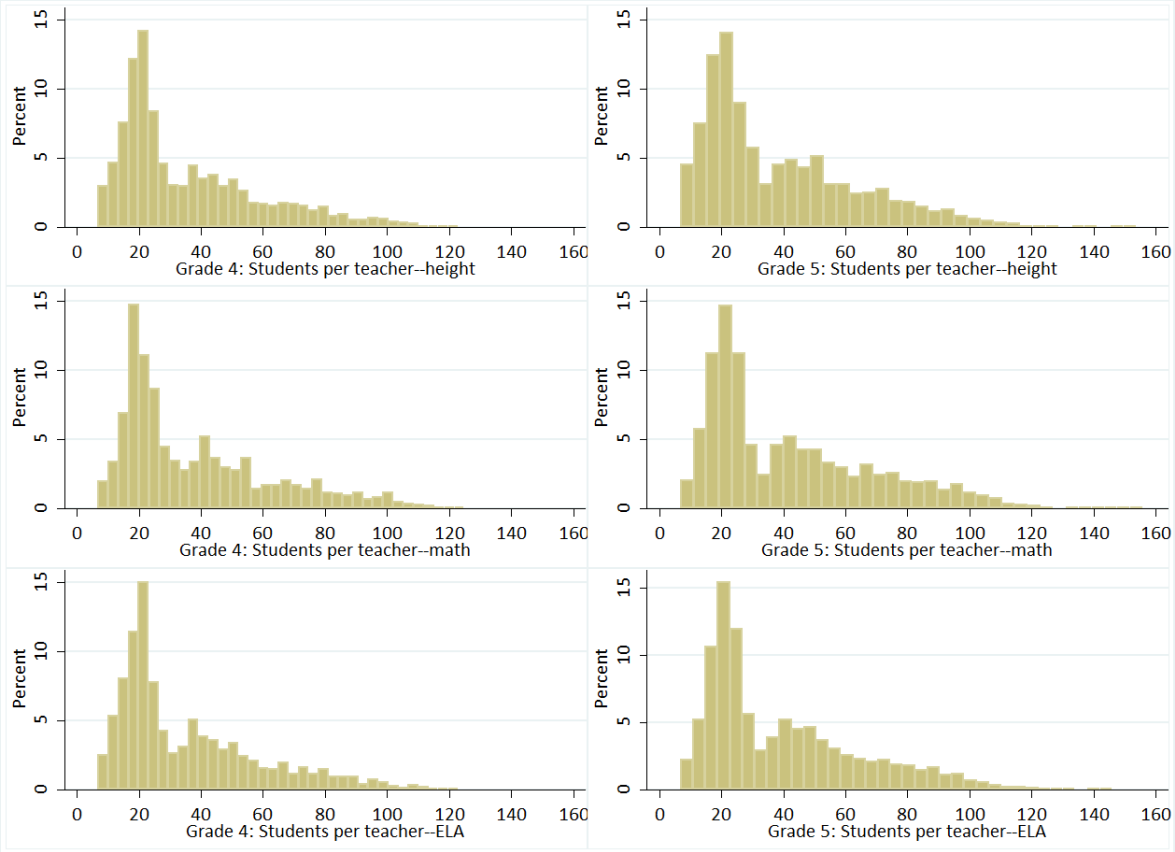
A Supplemental Appendix

This file contains supplemental appendix figures and tables. It is not intended for publication, but will be made available at the authors' websites.

The contents are as follows.

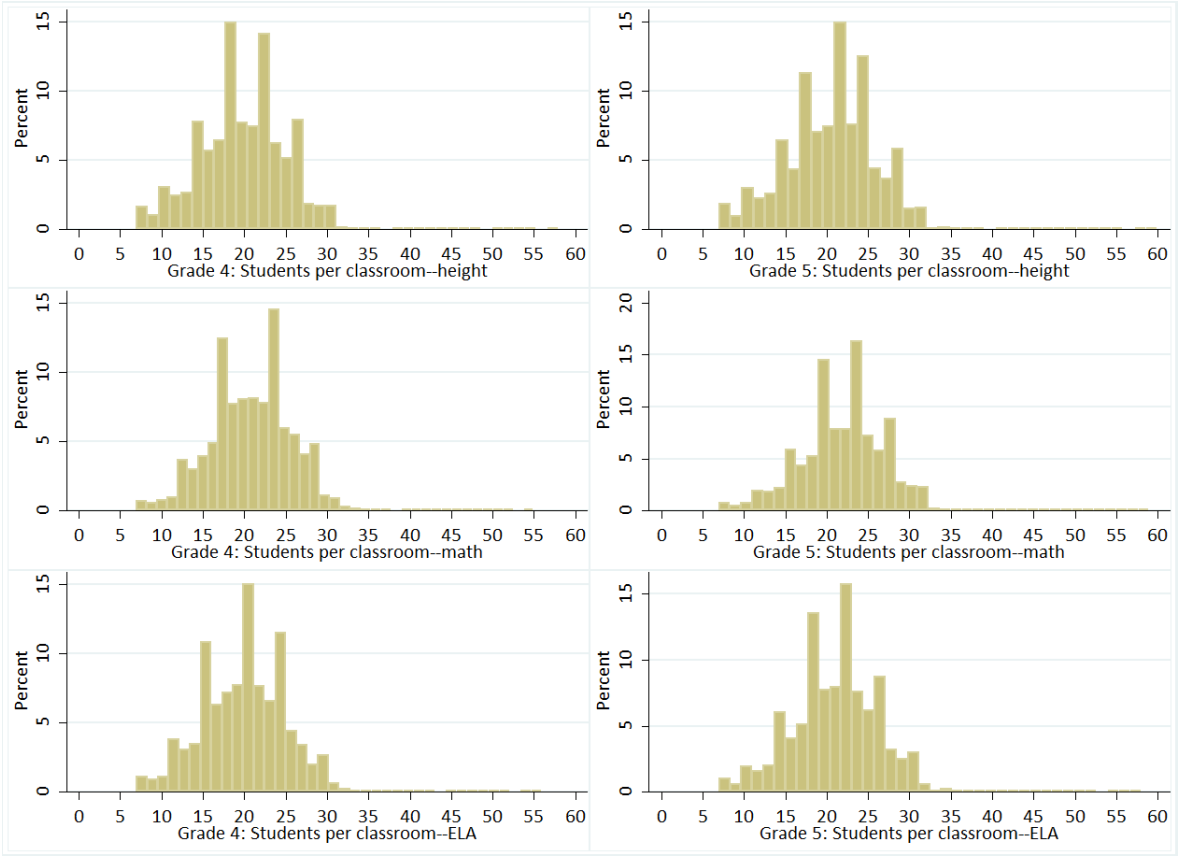
- Figure A.1 shows the distribution of students per teacher for our sample.
- Figure A.2 shows the distribution of students per classroom for our sample.
- Figure A.3 shows the distribution of estimated teacher effects on ELA achievement.
- Figure A.4 shows the distribution of p-values for tests of nonrandom tracking of students to classrooms by prior EA achievement.
- Table A.1 shows coefficients from regressions of achievement on height, lagged achievement, and the controls.
- Table A.2 shows coefficients from the baseline regressions used to calculate teacher value added on height.
- Table A.3 shows coefficients from the baseline regressions used to calculate teacher value added on math.
- Table A.4 shows coefficients from the baseline regressions used to calculate teacher value added on ELA.

Figure A.1: Students per teacher



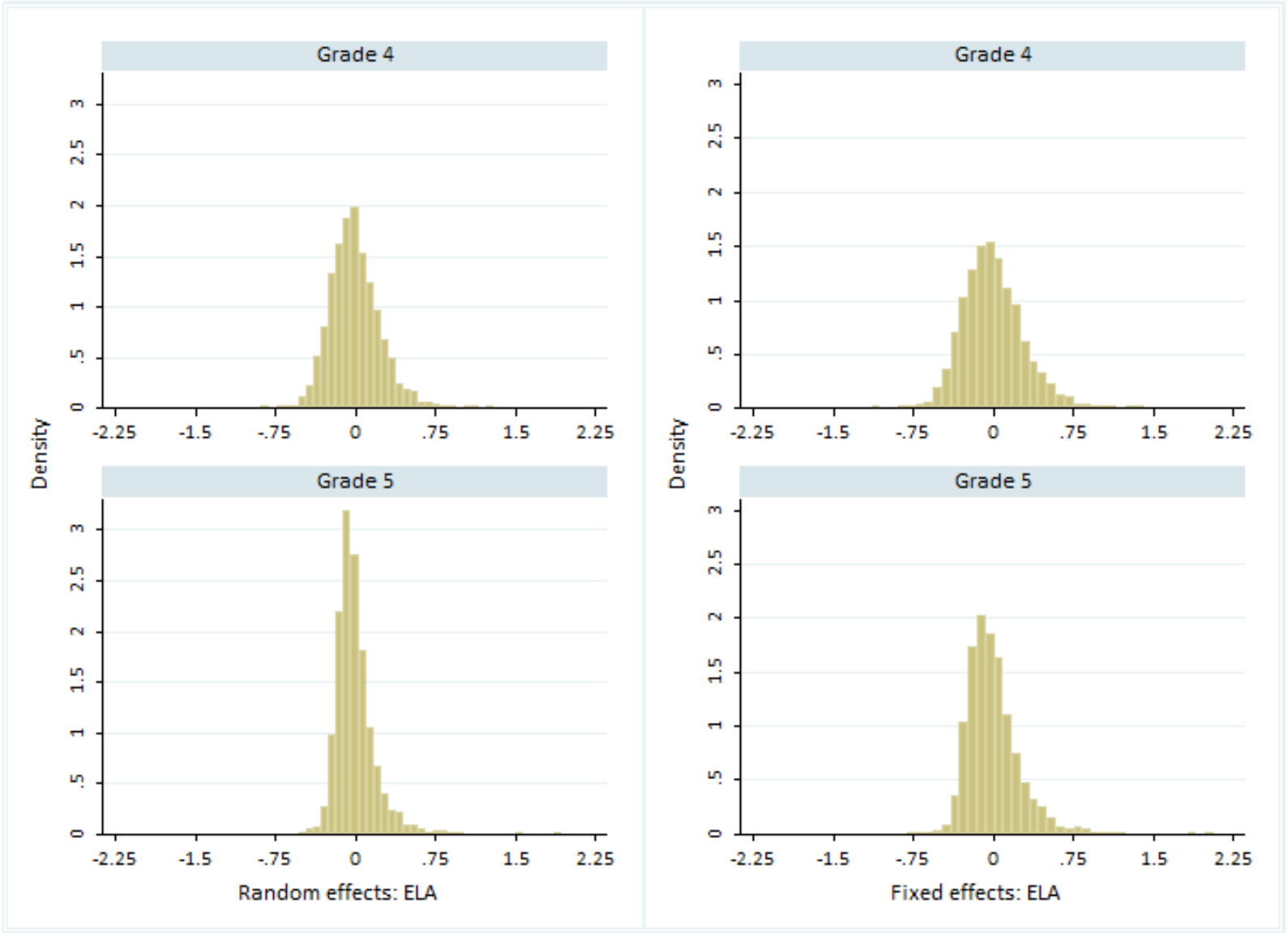
Notes: includes teachers with at least seven students with sufficient data for VAM models. A small number of teachers with more than 160 students over four years are excluded from the figure. For height, N=4261 and 3681. For math, N=4720 and 4242. For ELA, N=4366 and 3977.

Figure A.2: Students per classroom



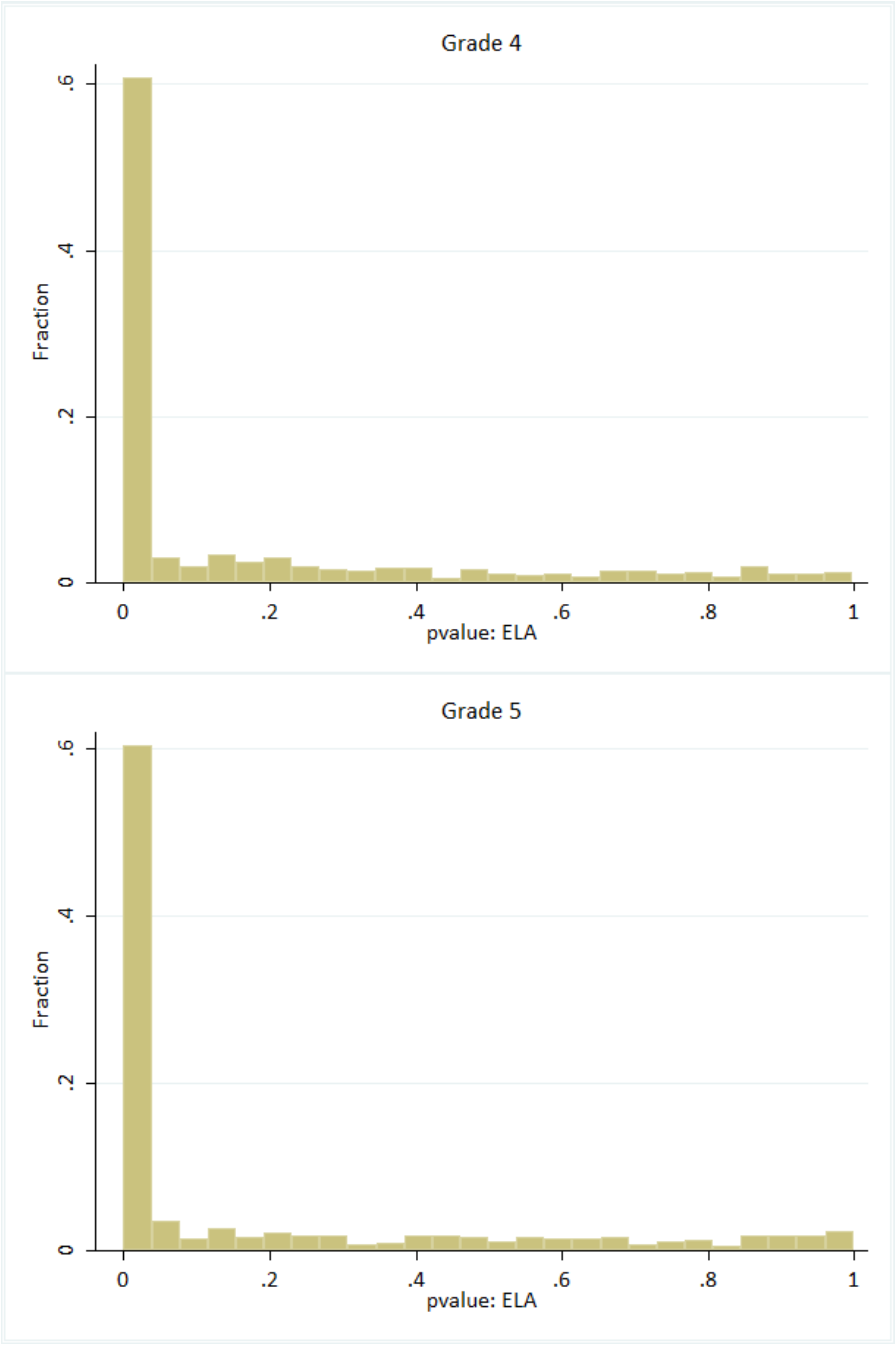
Notes: includes classrooms (teacher-years) with at least seven students with sufficient data for VAM models. A small number of classrooms with more than 60 students are excluded from the figure. For height, N=7590 and 6831. For math, N=8707 and 8109. For ELA, N=7788 and 7318.

Figure A.3: Distribution of teacher effects on ELA achievement



Notes: see notes to Table 4 for a description of how teacher effects were estimated.

Figure A.4: Tests for nonrandom tracking by prior ELA achievement



Notes: Each p-value is from a test of the hypothesis that classroom effects in a school s are jointly zero. Regression models are estimated separately for each school and grade, with lagged student outcomes regressed on school-grade-year and classroom effects.

Table A.1: Regressions of achievement on height

	ELA Grade 4	Math Grade 4	ELA Grade 4	Math Grade 4	ELA Grade 4	Math Grade 4	ELA Grade 5	Math Grade 5	ELA Grade 5	Math Grade 5	ELA Grade 5	Math Grade 5
Lag reading (z-score)	0.574***		0.574***		0.574***		0.578***		0.578***		0.584***	
	(0.002)		(0.002)		(0.002)		(0.002)		(0.002)		(0.003)	
Lag math (z-score)		0.595***		0.595***		0.591***		0.665***		0.665***		0.664***
		(0.002)		(0.002)		(0.002)		(0.002)		(0.002)		(0.002)
Height (z-score)	0.015***	0.012***	0.015***	0.012***	0.020***	0.017***	0.011***	0.011***	0.011***	0.011***	0.011***	0.013***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Height (z-score)$\hat{2}$			-0.004**	-0.003**					-0.006***	-0.003**		
			(0.001)	(0.001)					(0.001)	(0.001)		
Change in height z-score					-0.013***	-0.014***					-0.005	-0.004
(uses t-1 to t change)					(0.003)	(0.003)					(0.004)	(0.003)
Immigrant	0.085***	0.065***	0.085***	0.065***	0.086***	0.064***	0.066***	0.080***	0.066***	0.080***	0.065***	0.075***
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)	(0.007)	(0.005)
LEP	-0.335***	-0.272***	-0.335***	-0.272***	-0.346***	-0.291***	-0.129***	-0.177***	-0.129***	-0.177***	-0.128***	-0.185***
	(0.007)	(0.006)	(0.007)	(0.006)	(0.007)	(0.006)	(0.008)	(0.006)	(0.008)	(0.006)	(0.010)	(0.007)
Special education	-0.274***	-0.235***	-0.274***	-0.235***	-0.268***	-0.233***	-0.134***	-0.175***	-0.134***	-0.175***	-0.126***	-0.167***
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)	(0.007)	(0.006)
Low income	-0.176***	-0.162***	-0.176***	-0.162***	-0.177***	-0.162***	-0.145***	-0.075***	-0.145***	-0.075***	-0.137***	-0.079***
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)	(0.007)	(0.006)
Low income flag	-0.142***	-0.100***	-0.142***	-0.100***	-0.137***	-0.106***	-0.093***	-0.052***	-0.093***	-0.052***	-0.079***	-0.068***
	(0.009)	(0.008)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)	(0.007)	(0.009)	(0.007)	(0.011)	(0.008)
Bronx	-0.067***	-0.107***	-0.067***	-0.107***	-0.060***	-0.111***	-0.091***	-0.036***	-0.091***	-0.036***	-0.080***	-0.045***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.006)	(0.007)	(0.006)	(0.008)	(0.006)
Brooklyn	-0.026***	-0.048***	-0.025***	-0.048***	-0.026***	-0.058***	-0.042***	-0.006	-0.042***	-0.006	-0.040***	-0.015**
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)	(0.007)	(0.006)
Queens	0.017**	-0.048***	0.017**	-0.048***	0.016*	-0.062***	-0.037***	0.024***	-0.037***	0.024***	-0.043***	0.004
	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)	(0.007)	(0.006)
Staten Island	-0.092***	-0.159***	-0.092***	-0.159***	-0.094***	-0.169***	-0.065***	0.011	-0.065***	0.011	-0.063***	0.003
	(0.011)	(0.008)	(0.011)	(0.008)	(0.012)	(0.008)	(0.012)	(0.007)	(0.012)	(0.007)	(0.013)	(0.008)
Gender x race x age	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Year dummies	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Constant	0.625***	0.834***	0.625***	0.834***	0.549***	0.802***	0.770***	0.829***	0.772***	0.829***	0.702***	0.815***
	(0.145)	(0.129)	(0.145)	(0.129)	(0.157)	(0.139)	(0.177)	(0.134)	(0.177)	(0.134)	(0.195)	(0.146)
N	156605	182431	156605	182431	131544	153539	156842	180445	156842	180445	124286	144051

Notes: Each column reports coefficients from regressions of achievement on lagged achievement, height, lagged height, and the change in height, and demographics.

Table A.2: Coefficients from baseline regression models: height

	Height RE 4th grade	Height RE (SFE) 4th grade	Height FE 4th grade	Height RE 5th grade	Height RE (SFE) 5th grade	Height FE 5th grade
Lag height (z-score)	0.835***	0.826***	0.839***	0.842***	0.834***	0.845***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Immigrant	0.031***	0.031***	0.029***	0.003	0.005	0.001
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
LEP	-0.026***	-0.029***	-0.029***	-0.01	-0.015*	-0.011
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Special education	0.009	0.009	0.009	0.003	0.004	0.004
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Low income	0.007	0.007	0.006	0.011*	0.009	0.006
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Low income flag	-0.019*	-0.022**	-0.017*	0.001	-0.003	-0.003
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Bronx	0.023	-0.174	0.162*	0.040**	0.089	-0.016
	(0.013)	(0.096)	(0.076)	(0.014)	(0.145)	(0.083)
Brooklyn	-0.004	0.014	-0.233**	-0.007	0.168	-0.472***
	(0.012)	(0.097)	(0.080)	(0.013)	(0.137)	(0.130)
Queens	0.026*	0.008	-0.051	0.02	0.091	0.005
	(0.012)	(0.101)	(0.073)	(0.013)	(0.139)	(0.062)
Staten Island	-0.027	-0.255	-0.175	0.016	0.408	-0.680*
	(0.018)	(0.169)	(0.097)	(0.019)	(0.260)	(0.282)
Fitnessgram date diff	0.001***	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Gender x race x age	YES	YES	YES	YES	YES	YES
Year dummies	YES	YES	YES	YES	YES	YES
Constant	-1.171***	-1.140***	-1.053***	-1.847***	-1.942***	-1.621***
	(0.115)	(0.141)	(0.127)	(0.129)	(0.171)	(0.149)
N	153242	153018	153242	143738	143570	143738

Notes: Table reports baseline coefficients from estimation of value added for height. Columns 1–3 are for 4th graders and 4–6 for 5th graders. Columns 1–2, and 4–5 are random effects models and columns 3 and 6 are fixed effect models. Columns 2 and 5 add school fixed effects (SFE).

Table A.3: Coefficients from baseline regression models: math

	Math RE 4th grade	Math RE (SFE) 4th grade	Math FE 4th grade	Math RE 5th grade	Math RE (SFE) 5th grade	Math FE 5th grade
Lag math (z-score)	0.544***	0.571***	0.536***	0.632***	0.653***	0.626***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Immigrant	0.076***	0.073***	0.079***	0.081***	0.079***	0.082***
	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
LEP	-0.258***	-0.264***	-0.253***	-0.175***	-0.172***	-0.173***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Special education	-0.235***	-0.262***	-0.236***	-0.178***	-0.190***	-0.180***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Low income	-0.091***	-0.098***	-0.075***	-0.052***	-0.057***	-0.045***
	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Low income flag	-0.059***	-0.073***	-0.051***	-0.054***	-0.056***	-0.053***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Bronx	-0.118***	-0.145	-0.027	-0.055***	0.113	-0.077
	(0.015)	(0.124)	(0.069)	(0.014)	(0.136)	(0.061)
Brooklyn	-0.051***	-0.037	-0.435***	-0.018	-0.182	-0.07
	(0.014)	(0.125)	(0.081)	(0.013)	(0.152)	(0.083)
Queens	-0.017	0.069	-0.268***	0.041**	-0.017	-0.05
	(0.015)	(0.133)	(0.073)	(0.014)	(0.147)	(0.058)
Staten Island	-0.080***	0.176	-0.456***	0.027	-0.12	-0.303*
	(0.022)	(0.243)	(0.100)	(0.021)	(0.262)	(0.122)
Gender x race x age	YES	YES	YES	YES	YES	YES
Year dummies	YES	YES	YES	YES	YES	YES
Constant	0.677***	0.742***	0.873***	0.669***	0.791***	0.736***
	(0.122)	(0.160)	(0.133)	(0.127)	(0.171)	(0.137)
N	182623	182441	182623	180639	180576	180639

Notes: Table reports baseline coefficients from estimation of value added for math. Columns 1–3 are for 4th graders and 4–6 for 5th graders. Columns 1–2, and 4–5 are random effects models and columns 3 and 6 are fixed effect models. Columns 2 and 5 add school fixed effects (SFE).

Table A.4: Coefficients from baseline regression models: ELA

	ELA RE 4th grade	ELA RE (SFE) 4th grade	ELA FE 4th grade	ELA RE 5th grade	ELA RE (SFE) 5th grade	ELA FE 5th grade
Lag ELA (z-score)	0.507***	0.547***	0.491***	0.533***	0.559***	0.517***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Immigrant	0.096***	0.091***	0.098***	0.073***	0.071***	0.074***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
LEP	-0.339***	-0.336***	-0.332***	-0.145***	-0.122***	-0.148***
	(0.007)	(0.007)	(0.007)	(0.009)	(0.008)	(0.009)
Special education	-0.269***	-0.300***	-0.270***	-0.141***	-0.150***	-0.144***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Low income	-0.111***	-0.115***	-0.088***	-0.109***	-0.106***	-0.089***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
Low income flag	-0.111***	-0.124***	-0.101***	-0.064***	-0.069***	-0.051***
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
Bronx	-0.107***	-0.036	-0.147*	-0.103***	0.1	-0.081
	(0.014)	(0.133)	(0.073)	(0.013)	(0.168)	(0.077)
Brooklyn	-0.047***	-0.001	-0.241*	-0.053***	0.101	-0.236*
	(0.013)	(0.129)	(0.101)	(0.012)	(0.184)	(0.109)
Queens	0.040**	0.201	0.04	-0.021	0.111	0.025
	(0.014)	(0.139)	(0.099)	(0.013)	(0.177)	(0.087)
Staten Island	-0.009	0.205	-0.418	-0.021	0.182	-0.478
	(0.025)	(0.245)	(0.219)	(0.023)	(0.335)	(0.254)
Gender x race x age	YES	YES	YES	YES	YES	YES
Year dummies	YES	YES	YES	YES	YES	YES
Constant	0.411**	0.397*	0.474**	0.657***	0.603**	0.663***
	(0.139)	(0.174)	(0.153)	(0.172)	(0.217)	(0.186)
N	156767	156599	156767	157023	156967	157023

Notes: Table reports baseline coefficients from estimation of value added for ELA. Columns 1–3 are for 4th graders and 4–6 for 5th graders. Columns 1–2, and 4–5 are random effects models and columns 3 and 6 are fixed effect models. Columns 2 and 5 add school fixed effects (SFE).