Variability in the Grades Teachers' Give: Teacher Grade Value-Add, Mismatch, and Long-term Effects

Abstract

We identify systematic differences in ninth-grade teachers' effects on students' course grades in their class (ninth-grade grade value-add), their persistent effects on students' grades in the same subject the next grade (tenth-grade grade value-add, representing effectiveness at academic preparation), and mismatch between the grades they give and students' subsequent performance (representing inaccurate measurement). Both persistent effects and mismatch contribute to observed variation in grades. Teacher effects are stable across years, and their effects are predictive of students' long-term outcomes, including eleventh grade ACT scores, graduating GPA, and high school graduation. Persistent effects show consistent long-term benefits, while the relationship of grading mismatch with academic outcomes is nuanced.

Keywords: high school, value-added, grading practices, teacher effectiveness

Citation:

McKenzie, R. & Allensworth, E. (2025). Variability in the Grades Teachers Give: Teacher Grade Value-Add, Mismatch, and Long-Term Effects. *American Educational Research Journal l*: pp. 1–35. DOI: 10.3102/00028312251345094.

Introduction

Differences across teachers in the grades they give could matter considerably for students' educational outcomes. Student grade point averages are used for applications to programs, colleges, and scholarships. Grades also affect students' academic mindsets (e.g., self-efficacy, sense of belonging), influencing motivation and decision-making in future courses (Farrington et al, 2012). Students' grade point averages (GPAs) are strong predictors of student learning and post-secondary success (Atkinson & Geiser, 2009; Bastedo et al, 2023; Bowen, Chingos & McPherson, 2009; Bowers et al, 2013; Geiser & Santelices, 2007). In fact, Jackson (2018) found that ninth-grade teacher value-add on ninth-grade behaviors--including absences which are often a component of grading schema, and their value-add on tenth-grade GPAs–had larger impacts on students' long-term outcomes (high school graduation, college intentions, and final GPAs) than teachers' value-add on test scores.

However, teachers receive little-to-no objective feedback regarding the accuracy of the grades they assign, and there is considerable ambiguity for teacher evaluators in using data on teacher-assigned grades. One issue is uncertainty about why a teacher might give particularly low or high grades: is she especially effective/ineffective at motivating and teaching students so they legitimately earn those grades, or does she hold particularly high or low expectations of students, or grade in a harsh or lenient manner? A second issue is that people hold different beliefs about whether grades should be high or low: is it better to give higher grades to support students' self-efficacy or to give low grades so students are held to high standards? It is also not clear how much average course grades vary across teachers in real-world settings, and the extent to which grading differences across teachers matter for students in the long term.

This study presents a method for separating out ninth-grade teachers' effects on their students' grades in terms of their effectiveness at preparing students for coursework in the same subject the following year (their persistent effects), versus differences in their grading standards (their grading mismatch), and their total effects on their students' grades (preparation plus mismatch). It then examines the long-term impacts of having ninth-grade teachers who have different types of effects on students' grades.

Research on Grading Variability

Course grades are a multidimensional construct, representing teachers' assessment of students' skills, effort, persistence, and growth. Because they are assigned to students by individual teachers, students can receive different grades for the same quality of work based on differences in their teacher's standards, grading criteria, and severity/leniency (Brookhart et al 2016). Yet, differences in the average grades given by teachers could also result from differences in teacher effectiveness. A highly effective teacher could inspire students and foster strong work effort and learning, enhancing both behavioral engagement (attending class, completing assignments, discussing work with peers) and emotional engagement (enjoyment of learning, interest in subject), leading to better academic preparation and correspondingly high grades; likewise, an ineffective teacher may inspire little engagement in learning and foster lower grades (Fredricks, Blumenfeld, & Paris, 2004; Appleton, Christenson, & Furlong, 2008; Bransford, Brown, & Cocking, 2000; Mahatmya, Lohman, Matjasko, & Farb, 2018). Research on grading variability has often focused on issues of inconsistency, while not considering differences in effectiveness. As Brookhart et al (2016) note, research on grades "typically misses antecedent causes… for example, does … the variance in grades reflect achievement in classes where lessons are high-quality and appropriate for students?"

There are many studies of grading reliability showing differences across teachers in how they grade particular assignments, but little research that takes into account real-world settings, where teachers have different goals, expectations, and effectiveness. One large body of research has shown that teachers employ diverse rubrics and criteria for grading students, such that teachers can assign different grades for the same assignment, which is attenuated if they use the same grading rubrics (see Brookhart et al, 2016, for a thorough review). This research often occurs outside of teachers' actual classrooms, keeping the assignment constant. In the real world, teachers prioritize different skills, content, and behaviors, and their teaching and grading reflect these differences. Different teaching and grading priorities are not necessarily problematic; they could still lead to similar results in terms of students' academic preparation. What is more critical is whether the grades that teachers give differ in how well they represent students' academic readiness–students' academic skills and knowledge, their work effort, and their motivation and enthusiasm for the subject.

Another body of research has focused on the mismatch between grades and standardized test scores to assert that grades are unreliable (Camara, Kimmel, Scheuneman, Sawtell, 2004; Gershenson, 2018; Godfrey, 2011; Hurwitz & Lee, 2018). However, standardized tests are not designed to be comprehensive measures of academic performance in school. Grades capture cognitive skills and knowledge assessed through a broad variety of methods other than tests, as well as students' behavioral engagement, including attendance, effort, participation, assignment completion, and they also can reflect improvement relative to students' baseline skills (Brookhart et al, 2016; Kelly, 2008). Cognitive, non-cognitive, and behavioral outcomes are all important for students' long-term academic success (Cunha & Heckman, 2008; Jackson et al, 2024). As a result, we should not expect grades to completely align with test scores, and studies generally find correlations of about 0.5 between them (Brookhart et al, 2016; Pattison, Grodsky & Muller, 2013). We cannot rely on standardized test scores to identify grading unreliability among teachers because some of the mismatch may occur because teachers grade on factors that matter considerably for students but are not captured well by standardized tests.

Thus, it is not clear to what extent the grades that teachers give to students represent different levels of academic preparation. It is also not clear how much it matters for students that teachers assign different grades. Is it harmful to students if they have a teacher that tends to give particularly low grades? Or does it help them in the long run? As teachers make decisions about whether to be more lenient or stricter in their grading, they may hold conflicting perspectives about how to best foster students' subsequent outcomes. Teachers expect students to "earn" their grades through effort and performance, yet they also have concerns about the impacts on students' self-esteem and attitudes towards academic work in the subject (Brookhart, 1993). Students learn more when held to high expectations, especially when coupled with appropriate support to meet those expectations (Lee, 1999; Mitchell et al, 2015). At the same time students' mindsets about themselves as learners exert strong influence on subsequent motivation (Bandura, 1986; Caraway et al, 2003; Farrington et al, 2012); getting a low or high grade could harm perceptions of themselves as competent learners who can succeed in the subject area.

The limited research on the effects of grading practices on students' later outcomes suggests much is still not known. Two studies find that stricter grading practices lead to short term gains in student achievement (Mozenter, 2019; Betts & Grogger, 2003), but that the effects do not persist to influencing long-run outcomes (high school course selection, ACT scores, educational attainment), showing null or even negative effects. Grading strictness may have both positive and negative effects, which end up cancelling each other out, or depend on a teacher's overall effectiveness.

Research Questions and Conceptual Model

This paper attempts to isolate different components of grades given by ninth-grade teachers and discern their impacts on long-term student outcomes. As shown in Figure 1, we conceptualize teacher influence on student grades in their class in three ways. One is through their effectiveness at preparing students for the next class in their subject, which we call their persistent effects. This could occur through different mechanisms, such as influencing students' skill development, understanding of concepts, interest in the subject, or developing students' learning strategies. The second is through their idiosyncratic tendency to give grades that are higher or lower than other teachers give for similar levels of academic preparation (grade inflation or deflation), which we call grading mismatch. This could reflect different expectations of students' learning or work effort than is typical, and harsh or lenient grading. The combination of teacher effects on preparation with grading mismatch results in the total teacher effects on students' grades in their ninth-grade class, after controlling for the qualities students bring with them (e.g., prior grades and test scores, economic and demographic backgrounds) and characteristics of the class being taught (e.g., subject, level, prior achievement of students in the class), and the school (e.g., socioeconomic status of students, average eighth grade test scores).

We adapt the long- and short-run value-added approach used by Gilraine and Pope (2021) with test scores to the context of student grades to separate teachers' effects on academic preparation from their grading mismatch. The effects of ninth-grade teachers on students' academic preparation are determined through their students' performance in the fall of tenth grade in the same subject, holding constant students' achievement prior to ninth grade, background characteristics, and ninth- and tenth-grade course and school characteristics. These are the ninth-grade teacher's persistent effects, or their tenth-grade value-add.

We expect students with the same ninth-grade course grade to have similar fall tenth-grade grades in the corresponding subject area, holding constant characteristics of their ninth- and tenth-grade classes. The difference between a teacher's value-add on ninth-grade course grades and their persistent effects on their students' tenth-grade course grades is their grading mismatch. Mismatch identifies differences in measurement (grading inflation/deflation) while persistent effects identify differences in effectiveness.

This provides three teacher effects on student grades:

- Total effects on students' spring grades in the class taught by the teacher in ninth grade (i.e., the degree to which the grades the teacher gave were higher/lower than those of other teachers to similar students in similar classes, or their value-add on ninth-grade course grades);
- Persistent effects on students' grades in the same subject the following fall (i.e., their value-add on students' fall tenth-grade course grades); and
- Mismatch of the grades they gave in spring of ninth grade compared to their students' performance in the same subject in fall of tenth grade, with value-add adjustments (the difference between a teacher's total effect and their persistent effect).

We pose the following research questions:

1. To what extent do ninth-grade teachers have unique effects on students' course grades in the class they teach, so that students who have a particular teacher receive different course grades, on average, than students who have other teachers for the same course (holding constant student prior achievement and backgrounds, course and school characteristics)?

This indicates their value-add on students' ninth-grade course grades (total effects). These differences could exist because some teachers are more effective, so their students get higher grades because they have learned more than with other teachers. So we ask:

2. To what extent do ninth-grade teachers have persistent effects on their students' grades in the same subject the following year – so that students who have a particular ninth-grade teacher get higher or lower grades in the same subject in the fall of tenth grade, relative to students who had other teachers for that subject in ninth grade (holding constant student prior achievement and backgrounds, course and school characteristics)?

Differences in average grades also could exist because teachers give grades that are higher or lower than their students' academic readiness levels. So we ask:

3. To what extent do ninth-grade teachers consistently give grades that mismatch their students' performance in the same subject in the fall of tenth grade – so that their students perform better or worse the following year than would be expected based on the grades they received from the ninth-grade teacher (holding constant student prior achievement and backgrounds, course and school characteristics)?

To test whether persistent effects and mismatch are stable characteristics of individual teachers, and not solely defined by the students and courses they teach in a given year, we examine auto-correlations among the teacher effects over time, and students' ninth-grade outcomes based on their teacher's grade effects as calculated when they taught other cohorts of ninth-grade students. If grade effects are stable, we should see that students who are assigned a teacher that gave higher-than-average grades to other cohorts of students end up with higher grades in the year those students have the teacher, as well. We also examine spillover effects from having a particularly effective/ineffective or harsh/lenient ninth-grade teacher in one class on students' overall performance in ninth grade. We might expect, for example, that having a more effective teacher in one class (with positive persistent effects) might encourage stronger school attendance and higher grades overall, or that having a harsh grader (with negative mismatch) might make it harder for students to fulfill expectations in other classes. So we ask:

4. What is the relationship of students' ninth-grade grades, credits earned, and attendance with their teachers' total, persistent, and mismatch grade effects?

Finally, it is not clear whether it matters in the long term for students if their teachers give grades that are much higher or lower than the grades of other teachers, and whether it is better for teachers to give high grades or low ones, relative to other teachers. So we ask:

5. What are the long-term effects of having a teacher with large persistent grade effects or grading mismatch on the number of AP classes taken in tenth and eleventh grade, eleventh grade ACT score, graduating core GPA, and probability of graduating with a diploma within four years?

We estimate these effects for a sample of CPS teacher-year observations using a jackknife leave one-year out method so as not to confound the estimate of teachers' grade effects with the relationships of teacher grade effects to students' short- and long-term outcomes. We then calculate the impact of being assigned to teachers with large persistent effects, and with large grading mismatch, on a number of short- and long-term outcomes.

Data

The data for our analyses are drawn from longitudinal administrative records on students and teachers in the Chicago Public Schools (CPS). CPS teachers had autonomy in their grading practices during the years of the study. District guidelines published in 2017 noted that "teachers shall exercise their independent professional judgment in developing their grading practices.

They shall determine the number, type, weighting and frequency of student assignments and tests or other assessments that are used to determine individual course grades (CPS, 2017).” We focus on ninth-grade cohorts from the 2011-12 school year to the 2017-18 school year, and use data from school years 2010-2011 to 2018-19 in order to include information on students’ achievement prior to ninth grade, ninth-grade course information, tenth-grade course information, and long-term outcomes. We combine two primary sources of data on students:

Student level: The student-level dataset contains information on every student enrolled in the time period under consideration, with one observation per student per year. It contains: 1) demographic data (race and ethnicity, gender, IEP status, age), 2) academic records (test scores, graduation status, enrollment status), 3) post-secondary data from the National Student Clearinghouse providing college enrollment status, and 4) home address. Students’ addresses are tied to Census data at the block group level to create indicators of economic status based on the percent of families under the poverty line and male unemployment.

Student-Course-Grade level: The student-course-grade-level dataset provides student-by-course information on each of the courses a student took in a given semester, including the class subject and level (Honors, AP, regular, remedial), a unique teacher ID for the class, and the letter grade the student received. We convert the letter grades in this data to a 5-point numerical scale following the district’s convention, where an A is equivalent to four points, a B is equivalent to three points, and so on. We refer to this as grade-level data because each observation links a student to their grade in a single class.

The student-course-grade level dataset was also used to construct indicators of the composition of students in each course to control for peer compositional effects when calculating teacher effects. For each course, summary variables were created based on aggregates of student data that include: number of students in the class, class mean math and reading eighth grade test scores (linear, quadratic and cubic terms), class mean age, class mean eighth grade attendance, class mean eighth grade GPA, percentage of students with IEPs, and average percent of families below the poverty line in students' residential census blocks. Corresponding school-level variables were also constructed for each semester.

Together, these controls help us accurately construct teacher grade effects by comparing outcomes for students in the same type of course and subjected to similar peer effects. Studies that create value-add indicators at the high school level must be particularly careful about issues of selection of students into different types of courses. In this case, the information we have on the characteristics of the students in each class, and on individual students, is so extensive that we feel confident in producing our estimates of teacher effects, discussed further below under the estimation procedure for total effects. Still, while our estimates here should hold true on average, the grade effects estimated for any individual teacher could be affected by course selection or other unmeasured patterns.

Sample Construction

From the two datasets given above, we construct our primary dataset that contains one observation per student-subject pair in each ninth-grade cohort. This serves as the main analytic sample for the estimation of our teacher grade effects, and we use it to calculate the relationships of teacher effects with students' short-term outcomes.

Each record contains data on the student's grade in the subject in ninth grade, and their grade in the subject in tenth grade, as well as information about the student's background (e.g., demographic information, eighth grade grades and test scores), and information about the course (e.g., level of course, average eighth grade achieve. We limit our sample to students who were ninth-graders in school year 2011-12 to 2017-18 who: attended a CPS high school excluding charter or alternative schools as their transcripts are not available through central administration files (but including selective and specialty schools); did not leave the school district during this year; have non-missing observations for the controls used in our estimation of teacher grade effects (see “Estimation Procedure” for control variable lists).

We further limit our sample to one subject-by-student-by-semester observation for each of the four core subjects (English, math, science, and social studies). For students enrolled in two classes of one subject (e.g., Algebra I and another math elective), we select the class taken more often district-wide (e.g., Algebra I). In the rare cases when more than one teacher is listed for a ninth-grade class, we drop the observation. We do not include: observations with missing final grades and those from classes not taken for credit; classes taught by teachers who taught less than seven students in a given semester; classes that contained over 25 percent special education students; ninth-grade classes without a valid and unique teacher ID; and the 2015-16 freshman cohort due to a temporary incomplete coverage of teacher IDs in the data for this year (accompanying a switch to a new data system).

To calculate the relationships of teacher effects with students’ long-term outcomes, we use a sub-sample that does not include the 2017-18 freshman cohort which was in ninth grade too recently to have complete data on long-term outcomes. We also drop from this sample students who left CPS before their senior year for a reason other than graduating or dropping out (such as a validated transfer or institutionalization).

Table 1 provides the number of students in each ninth-grade class in the district unfiltered, as well as the number of students and demographic characteristics in the full analytic sample and the long-term outcome sub-sample. The analytic sample is about half the size of total population as it does not include students in alternative or charter high schools (about a quarter of the total), and students who did not attend CPS in eighth grade or do not have eighth grade GPAs because they attended a charter elementary school (about a third of the total), and students who did not remain through fall of tenth grade. The analytic sample has a smaller percentage of Black students (32 percent versus 41 percent) than the total ninth-grade population. This is consistent with racial and ethnic differences in charter school enrollment (see Gwynne & Moore, 2017). It is also slightly more female (52 percent versus 50 percent).

Estimation Procedure for Teacher Effects

We develop estimates of three types of ninth-grade teacher grade effects: total effects, persistent effects, and grade mismatch, calculated separately for each core subject: mathematics, English, science, and social science.

Total Grade Effects (Ninth-Grade Value-Added)

We begin with an estimation of ninth-grade teachers' value-add on students' spring grades when in their class. Their "total grade effect" is an estimate of how much higher or lower students' grades are with that teacher, on average, relative to an average ninth-grade teacher, controlling for their students' eighth grade achievement, backgrounds, and course characteristics. Students get higher grades than expected if they have a teacher with a positive total grade effect and lower grades if they have a teacher with a negative total grade effect.

We use a jack-knife leave-one-year-out approach, described below, so that a teacher's grade effect in a given year is based on the grades she gives in every other year (with different students). In this way, when we match teacher grade effects to students to study student outcomes, the grade effects are not estimated with the same group of students with which we measure student outcomes. It ensures that we are measuring a characteristic of the teacher and not the specific conditions of a class in a given year. Note that teachers must have observations in at least two years to be included.

To estimate the total grade effect, we closely follow Gilraine and Pope (2021) to adapt the two-step value-added estimation method commonly used in the teacher value-added literature (Chetty, 2014). We define the total grade effect, γ_{jt} , as the amount which teacher j on average increases the grade of their students during year t. Therefore, the spring ninth-grade course grade, ${g_{ijt}}^*$, of student i taught by teacher j during year t (their ninth-grade year), is given by:

$${g_{ijt}}^* = \beta X_{ijt} + \gamma_{jt} + U_{ijt}$$

where X_{ijt} are observable determinants of the student's grade (student- and class-level controls), γ_{jt} is teacher's j's total grade effect in year t, and U_{ijt} represents residual idiosyncratic student variation.

In their study of high school value-add on test scores, Goldhaber, Goldschmidt and Tseng (2013) found that including student fixed-effects improved the estimation of teacher effects beyond just using a lagged score. We decided not to include student fixed effects when calculating teacher value-add because we thought it likely that a high (or low) value-add teacher on grades could influence the grades students get in their other classes. Including student fixed-effects would mask this phenomenon. Instead, we include a very large array of student- and class-control variables to account for potential sorting of students to particular teachers that is much more extensive than lagged achievement alone. These not only include a series of lagged achievement variables, but also the number of Honors and AP classes taken as a ninth-grader (capturing a student's tendency to enroll in more demanding classes), and a series of classroom compositional variables that capture the characteristics of students that enroll in that particular class (see Kane et al, 2013; Kane & Staiger, 2008; Jackson, 2018; Protic et al 2013).

The full list of control variables includes year fixed effects and the following student-level variables:

- 8th grade math score (standardized within year)
- 8th grade reading score (standardized with year)
- 8th grade attendance (in percentage terms)
- 8th grade core GPA
- SPED Status
- Poverty level in the student’s home census block (constructed from the percentage of households below the poverty line and male unemployment rate)
- Race, Ethnicity, and Gender Indicator Variables
- Student Age in ninth grade
- Number of Honors classes taken in ninth grade
- Number of AP classes taken in ninth grade

Ninth-grade course-level and school-level controls (with separate variables for course and school) include:

- Course level (indicators for honors vs AP vs regular)
- Course and School size (number of students)
- Course and School average standardized 8th grade math score (up to cubic term)
- Course and School average standardized 8th grade reading score (up to cubic term)
- Course and School average 8th grade attendance
- Course and School average 8th grade core GPA
- Course and School percent with IEPs
- Course and School average poverty (students’ residential block groups)

We elected not to include school fixed-effects in the models so as not to discount school organizational factors that might lead teachers to be more/less effective or more/less consistent in their grading than teachers at other schools who serve similar students in similar classes. A school that has a strong teacher mentoring program, for example, might have more effective teachers than a school that does not, leading students to get higher grades because the quality of teaching in their school is better than at other schools. As a result, there may be school effects that are uncorrelated with average student incoming achievement, attendance, poverty, and school size incorporated into the estimates of teacher effects. We discuss this further later.

We follow the teacher value-added literature and work with a residualized version of each student's spring course grade (g_{ijt}), with the impact of all observable predictors of this grade removed:

$$g_{ijt} \equiv g^*_{ijt} - \beta X_{ijt}$$

where β is estimated using the following OLS regression:

$$g^*{}_{ijt} = \alpha_j + \beta X_{ijt} + \epsilon_{ijt}$$

We include teacher fixed effects (α_j) when estimating β so that any component of the value-added which is correlated with elements of X_{ijt} is not mistakenly attributed to X_{ijt}. This estimates the relationships of $X_{ijt}$with students' grade net of any potential teacher effects. We use only a student's spring semester grade for $g^*{}_{ijt}$, to capture the impact of an entire year with their ninth-grade teacher on their grades. Models are run separately for each subject area.

The residualized grade is then a product of the teacher value-add (their total grade effect) plus residual idiosyncratic student variation not attributable to student background or course characteristics:

$$g_{ijt} \equiv g^*_{ijt} - \beta X_{ijt} = \gamma_{jt} + U_{ijt}$$

While we are able to see here that $\gamma_{jt} + U_{ijt} \ = \ \alpha_j + \epsilon_{ijt},$ it is important to note that it does not necessarily hold that $\alpha_j \ = \ \gamma_{jt}$. This is because α_j merely captures the residualized means of grades a teacher gives (which may be impacted by idiosyncratic shocks), while γ_{jt} represents their "true" (and unobservable) impact on student grades, which we try to estimate.

To estimate γ_{jt} in the equation above, we follow Chetty (2014) and use a jack-knife empirical Bayes estimator. Specifically, we let

$$\bar{G}_{jt} = \frac{1}{n} \sum_{\mathrm{i} \in \{\mathrm{i:j(it)=j}\}} g_{it}$$

denote the mean (residualized) grade in the ninth-grade class taught by teacher j in year t. Let

$$\bar{G}_j^{-t} \equiv (\bar{G}_{j1}, \bar{G}_{j21}, \dots, \bar{G}_{j,t-1}, \bar{G}_{j,t-2}, \dots, \bar{G}_{jT},)'$$

denote the vector of mean (residualized) grades in the ninth-grade classes taught by teacher j in every year besides t. We then use the following OLS regression to obtain the best linear predictor of $\bar{G}_{jt}$, the missing year from the vector above:

$$\overline{G}_{jt} = \psi \overline{\mathbf{G}}_j^{-t}$$

where:

$$E[\psi] = \frac{Cov(\overline{G}_{jt}, \overline{\mathbf{G}}_j^{-t})}{Var(\overline{\mathbf{G}}_j^{-t})}$$

Using the empirical analogue for ψ, teacher j's total grade effect in year t is given by:

$$\hat{\gamma}_{jt} = \hat{\psi} \overline{\mathbf{G}}_j^{-t}$$

For $\hat{\gamma}_{jt}$ to represent an unbiased estimate of teacher j's total grade effect, we need the following assumption:

Assumption 1 (Random Assignment of Teacher j(it)): Let $j(it)$ represent the teacher assigned to student i during their ninth-grade year, t. Then we assume that our set of residualizing control variables X_{ijt} are together comprehensive enough predictors of student's grades that the remaining unobservable determinants of students' grades are uncorrelated with the total grade effects of the teachers students are assigned to:

$$E[U_{ijt} | \, j(it)] = E[U_{ijt}]$$

In other words, students should not be sorted into high- or low-total grade effect teachers based off any unobserved predictor of their grades in ninth grade. We think this assumption is likely met because we control for a very large array of eighth grade achievement variables, a variable representing student proclivity for taking high-level courses (number of Honors/AP courses in ninth grade), as well as variables representing the peer composition of students in a class.

Persistent Grade Effects – Tenth-Grade Value-Added of Ninth-Grade Teachers

We next adopt a similar approach to estimate a ninth-grade teacher's value-added on their students' grades in the fall of their tenth-grade year - the semester after they were assigned to the teacher in question. We call this their "persistent grade effect". It estimates how much higher or lower students' sophomore fall semester grades are if they had that ninth-grade teacher than with the average teacher.

Intuitively, this value-added measure represents the component of the teacher's impact on student learning which persists into their students' grades the following term. This is conceptually similar to Jackson's (2018) prediction of ninth-grade teachers' value-add on their students' tenth-grade GPA but differs by calculating teachers' effects in the same subject, and just on fall term, rather than their effects on total tenth-grade GPA. We focus on grades only in the fall term to minimize tenth-grade teacher effects. In addition, for practical application, focusing on fall grades could allow data on teachers' persistent effects to be constructed and given to teachers while they are teaching their next year's class, allowing for timely feedback for instructional improvement.

Using a similar set-up to the previous section, we can note that student i's tenth-grade fall semester grade when assigned to teacher k in period $t + 1$ is given by:

$$g_{ik,t+1}{}^{*} = \beta_1 X_{ijt} + \beta_2 X_{ik,t+1} + \tau_{jt} + U_{ik,t+1}$$

where $g_{ik,t+1}{}^{*}$ is the student's grade, τ_{jt} is teacher j's (their ninth-grade year teacher) persistent grade effect, X_{ijt} contain all of the same predictors as the total effects (student eighth grade achievement, background, ninth-grade course composition), plus $X_{ik,t+1}$includes additional controls for course composition from the tenth-grade year:

- 10th grade course level
- 10th grade class size
- Average eighth grade ELA score of students in 10th grade class
- Average eighth grade mathematics score of students in 10th grade class
- Average eighth grade GPA of students in the 10th grade class
- Average poverty level of students in the 10th grade class
- Average age of students in the 10th grade class

$U_{ik,t+1}$ contains any remaining factors which could impact the student's tenth-grade grade. Then, as before, residualized tenth-grade year grades $g_{ik,t+1}$are given by:

$$g_{ik,t+1} = \tau_{jt} + U_{ik,t+1}$$

We can decompose the error term of the above equation, $U_{ik,t+1}$, as

$$U_{ik,t+1} = e_{i,t+1} + u_{ik,t+1}$$

where $e_{i,t+1}$ are non-observable *student-level* components of a student's tenth-grade year grade, and $u_{ik,t+1}$ are other non-observable *teacher-level* components of a student's tenth-grade year grade. For example, $u_{ik,t+1}$ could contain the portion of the student's ninth-grade teacher's grading practices that which are not accounted for by the observable characteristics in X_{ijt} and $X_{ik,t+1}$.

Using this decomposition, we can write two additional assumptions necessary for the identification of τ_{jt} in the equation above:

Assumption 1a (Random Assignment of Teacher j(it)): *Let* $j(it)$ *represent the teacher assigned to student* i *during their ninth-grade year,* t. Then we assume that students are not sorted into their ninth-grade teachers according to unobservable predictors of their tenth-grade grade:

$$E[e_{i,t+1} | \, j(it)] \; = \; E[e_{i,t+1}]$$

This extends Assumption 1 by additionally assuming that the observed student-level predictors of a student's tenth-grade grade are sufficiently rich so that the remaining unobserved student-level predictors of their tenth-grade grade are uncorrelated with the grading effect of the teacher they are assigned to in ninth grade.

Assumption 2 (Random Assignment of Teacher $\boldsymbol{k(i,t+1)}$): *Let $k(i,t+1)$ represent the teacher assigned to student i during their tenth-grade year, $t+1$.* As stated above, we let $u_{ik,t+1}$ denote the non-observable teacher-level components of student i's grade. Then we assume:

$$E[u_{i,t+1} | \, j(it)] \; = \; E[u_{i,t+1}]$$

This assumes that the effects of a student's teacher in their tenth-grade course on grades are uncorrelated with their ninth-grade year teacher's effects, beyond the observable characteristics included as controls in the model.

Under these assumptions, estimation of τ_{jt} is very similar to the estimation of the ninth-grade value-added described above, but replacing $\bar{G}_{jt}$ with:

$$\bar{T}_{jt} = \frac{1}{n} \sum_{\mathrm{i} \in \{\mathrm{i:j(it)=j}\}} g_{ik,t+1}$$

Then, letting

$$\bar{T}_j^{-t} \equiv (\bar{T}_{j1}, \bar{T}_{j21}, \dots, \bar{T}_{j,t-1}, \bar{T}_{j,t-2}, \dots, \bar{T}_{jT},)'$$

denote the vector of mean (residualized) tenth-grade course grades of the ninth-grade classes taught by teacher j in every year besides t, we can use OLS to obtain the best linear predictor of $\bar{T}_{jt}$, the missing year from vector above: $\bar{T}_{jt} = \varphi \overline{\mathbf{T}}_j^{-t}$

Using the empirical analogue for φ, teacher j's persistent grade effect in year t is given by:

$$\hat{\tau}_{jt} = \hat{\varphi}\overline{\mathbf{T}}_{j}^{-t}$$

Grading Mismatch

Grading mismatch captures differences between the grades a teachers' students receive in spring of their ninth-grade year (with that teacher) and the grades they earn in fall of their tenth-grade year in the same subject, with value-added adjustments. Imagine, for example, that students in the district have an average of 3.0 in both their ninth- and tenth-grade years. If a particular teacher's students average a 3.5 in their ninth-grade year and still go on to have a 3.0 average during their tenth-grade year, then it would appear that this teacher gives grades that systematically overestimate their student's academic readiness in the subject, compared to the district average. Similarly, if a ninth-grade teacher's students average a 3.0 in their class but go on to average a 3.5 in the tenth grade, then that teacher could be seen to systematically underestimate the preparation level of the students in their class.

To estimate mismatch, we follow the same value-added methodology discussed above,

$$m_{ijt}{}^{*} = g_{ijt}{}^{*} - g_{ik,t+1}{}^{*} = \beta_1 X_{ijt} + \beta_2 X_{ik,t+1} + \mu_{jt} + U_{ijt} + U_{ik,t+1}$$

where $m_{ijt}{}^{*}$ is the raw difference in grades between student's ninth-grade spring and tenth-grade fall grades, μ_{jt} is teacher j 's grading mismatch effect, X_{ijt} and $X_{ik,t+1}$ contain observable predictors of a student's tenth-grade year grade from both their ninth- and tenth-grade years, and U_{ijt} and $U_{ik,t+1}$ contains any remaining factors which could impact the student's ninth- and tenth-grade course grades. From here, we get the residualized version of the mismatch:

$$m_{ijt} = \mu_{jt} + U_{ijt} + U_{ik,t+1}$$

where μ_{jt} is still identified under our assumptions from above. From there, our estimation proceeds identically to that of the ninth- and tenth-grade value-added measures, but replacing $\bar{G}_{jt}$ with:

$$\overline{M}_{jt} = \frac{1}{n} \sum_{i \in \{i:j(it)=j\}} m_{ijt}$$

Then, letting $\bar{M}_j^{-t} \equiv (\bar{M}_{j1}, \bar{M}_{j21}, \dots, \bar{M}_{j,t-1}, \bar{M}_{j,t-2}, \dots, \bar{M}_{jT},)'$

denote the vector of mean (residualized) tenth-grade grades of the ninth-grade classes taught by teacher j in every year besides t, we can use OLS to obtain the best linear predictor of $\bar{M}_{jt}$ the missing year from vector above:

$$\bar{M}_{jt} = \Phi \bar{M}_j^{-t}$$

Using the empirical analogue for φ, teacher j's persistent grade effect in year t is given by:

$$\hat{\mu}_{jt} = \widehat{\Phi} \bar{M}_j^{-t}$$

Limitations

We assume that a ninth-grade teacher's students are randomly assigned to tenth-grade teachers, net of covariates. On average, students from a given ninth-grade teacher enrolled the following year in courses taught by five different tenth-grade teachers in their subject, with a median of four tenth-grade teachers. However, any systematic sorting of a ninth-grade teacher's students to tenth-grade teachers who themselves have grade effects will be incorporated into the ninth-grade teacher's persistent and mismatch effect. This issue is discussed further below and in the online technical appendix.

Mismatch: This approach assumes that the effectiveness of ninth-grade teachers at improving their students' knowledge and skills should be visible in their students' performance in the same subject the following year. To the extent that teachers in a given subject are teaching generalizable skills and attitudes (e.g., scientific thinking, analysis of texts, interest in history), we would expect that students with more effective ninth-grade teachers show stronger performance the following year. However, a teacher's students might receive grades that are idiosyncratically high when compared with their tenth-grade performance because the teacher employs a lenient grading policy, *or* because that teacher creates temporary increases to their students' skills that do not persist to their tenth-grade year. The methodology we employ does not allow us to differentiate between these possible explanations. For ease of describing mismatch in an intuitive way, we refer to teachers with positive mismatch as those that are more lenient, and those with negative mismatch as harsher, but recognize that they may simply grade on factors not valued in tenth-grade classes.

Persistent Effects: We also purposefully do not include school fixed effects to allow for cases in which teachers may be working as a team, for example, improving the effectiveness of all members of a particular department. This means, however, that any potential school effects on student outcomes not accounted for by control variables are incorporated in our estimate of teacher effects. Estimates controlling for school effects are discussed in the technical appendix.

Analysis of Teacher Grading Effects on Student Outcomes

To estimate the impact of each type of teacher grade effect on student outcomes, we regress a residualized version of the student outcome on the teacher effect of their ninth-grade teacher, pooling all grades and years. Student outcomes are residualized using the same vector of observable demographic and prior achievement characteristics used in residualizing their ninth- and tenth-grade grades for the estimation of the teacher effects. In this way, we control for all covariates prior to estimating teacher effects, which attributes any shared variance that might exist between the covariates and the teacher effects solely to the covariates. As before, if ${y_{ijt}}^*$

represents the raw value the outcome y for student i assigned to teacher j during their ninth-grade year t, then the residualized version of this outcome is given by:

$$y_{ijt} \equiv y_{ijt}{}^{*} - \beta X_{ijt}$$

where X_{ijt} includes the same vector of controls used to residualize grades in the prior section. For ninth-grade student outcomes we use the version of X_{ijt} that does not include tenth-grade class- and school-level controls. For the long-term student outcomes (number of AP courses after ninth grade, ACT score, graduating core GPA, probability of graduating in four years) we add to X_{ijt} the additional tenth-grade course- and school-level controls described above in models of the estimates of teacher persistent effects. As before, the coefficient β is estimated using an OLS regression of the form:

$$y^{*}{}_{ijt} = \alpha_j + \beta X_{ijt}$$

Once we have these residualized outcomes, we use OLS to estimate the equation:

$$y_{ijt} = \pi_1 \hat{\gamma}_{jt} + \epsilon_{ijt}$$

where $\hat{\gamma}_{jt}$ is the estimated total grade effect of the teacher j assigned to student i in period t and ϵ_{ijt} represents other idiosyncratic variation in student grades. When estimating this regression, this error term is clustered at the school and year level. We additionally estimate:

$$y_{ijt} = \pi_2 \hat{\tau}_{jt} + \pi_3 \hat{\mu}_{jt} + \epsilon_{ijt}$$

where, as before, $\hat{\tau}_{jt}$ and $\hat{\mu}_{jt}$ were the estimated persistent and mismatch grade effects of the teacher j assigned to student i in period t respectively. We group these effects in a single specification to observe the independent impact of each on student outcomes while the other type of effect is held constant. As above, ϵ_{ijt} represents other idiosyncratic variation in student grades and is clustered at the school and year level.

Results

Distribution of Grade Effects and their Intercorrelations

Summary statistics for teachers' total, persistent, and mismatch grade effects are given in Table 2. By design an average ninth-grade teacher has a grade effect of zero, while grades are higher when ninth-grade teachers have positive effects and lower when ninth-grade teachers have negative effects.

The standard deviations show the degree of variation across teachers in grade effects and mismatch. We take two approaches to the standard deviations, both shown in Table 2. The first is the empirical standard deviation of the estimated grade effects. Because the grade effects are fitted towards zero, this standard deviation will underestimate the true deviation of teacher grade effects in practice. To attempt to address this, we follow Chetty (2014) and also estimate the standard deviation of each grade effect as the square root of the within-year covariance of residuals across a random pair of classrooms taught by the same teacher. If we consider that two times the standard deviation should approximate the second and 98th percentiles (in a normal distribution), the alternative standard deviations better match the observed minimum and maximum than the empirical standard deviations. In fact, twice the alternative standard deviation is generally close to the observed minimum and maximum grade effects across the different subjects and types of effects estimated. At the same time, twice the alternative standard deviation is sometimes more than the observed minimum and maximum, particularly for estimates of mismatch, and we would expect some outliers with the minimum and maximum. Thus, the alternative standard deviations may be slight over-estimates of the true standard deviations, while the empirical standard deviations are under-estimates.

A ninth-grade teacher with a total grading effect one standard deviation above or below the mean gives grades that are, on average, 0.379 - 0.421 of a letter grade different than an average ninth-grade teacher gives to students with similar incoming achievement and demographics in similar classes, using the adjusted standard deviations. Ninth-grade teachers two standard deviations from the mean, or about the 2nd and 98th percentiles, give grades that are about 0.8 of a GPA point (between 0.758 - 0.841) higher or lower than average ninth-grade teachers give to similar students in similar courses. The standard deviations of the persistent effects (0.348 - 0.374) show that students' grades are higher or lower in fall of tenth grade by over a third of a GPA point for teachers one standard deviation higher or lower than average, and about 0.7 GPA points (0.696 - 0.748) for teachers two standard deviations higher or lower than the average ninth-grade teacher. The standard deviation of mismatch (0.423 - 0.469) suggests that some teachers' ninth-grade grades are higher or lower than typical given their students' subsequent performance by just under a half of a GPA point if they are one standard deviation above/below average, or just under one GPA point (0.846 - 0.938) for teachers at the extremes (two standard deviations above or below the mean).

The standard deviations in ninth-grade teacher effects are similar across subjects. There is more variance in mismatch among ninth-grade science teachers than teachers of other subjects, which could reflect greater substantive differences between ninth- and tenth-grade science classes than between ninth- and tenth-grade classes in other subjects. At the same time, high variance in persistent effects in science suggests that which ninth-grade teacher a student has for science matters in substantive ways for their tenth-grade science class performance. The largest variance in persistent effects occurs with ninth-grade science and social studies teachers, suggesting which teacher a student has for ninth-grade science or social science matters slightly more for their tenth-grade class than which teacher they have for English or math. Because of similarities across subjects in teacher effects, including in the relationships of teacher effects with later outcomes (not shown here), we present pooled results in the subsequent tables that examine teacher grade effects on student outcomes (entering observations for all four subjects simultaneously in the models).

Ninth-grade teachers' total effects are about equal to the sum of their persistent effect plus their mismatch, although there are some cases where they differ because each was estimated separately. The variance of the total grade effect is not necessarily wider than that of the persistent grade effects or mismatch because there is a negative correlation of -0.482 between persistent effects and mismatch. Ninth-grade teachers with positive mismatch (e.g., lenient graders) are less likely to have persistent positive effects on students' grades than ninth-grade teachers with negative mismatch (harsh graders). At the same time, Figure 2 shows that there are ninth-grade teachers with every combination of high/low persistent effects and positive/negative mismatch.

Total effects are more strongly correlated with mismatch than persistent effects, with correlations of 0.603 and 0.360, respectively. Thus, total grade effects are more likely to reflect grading mismatch (poor measurement) than persistent effects of ninth-grade teachers on achievement. At the same time, the positive correlation between total and persistent effects (0.360) indicates that ninth-grade teachers with positive persistent effects tend to give higher average grades than other ninth-grade teachers. It is just that their grades tend not to be as high as their students' later performance suggests they should have been, so many have negative mismatch. The corollary is also true: ninth-grade teachers with negative persistent effects tend to give lower than average grades, but the grades they give are not as low as their students' later performance would suggest they should have been, so many show positive mismatch (leniency).

Ninth-grade teachers' grade effects tend to be stable over time, with auto-correlations across years that are similar or higher than those reported for value-added based on test scores, which tend to range from 0.18 to 0.64 (Koedel & Rockoff, 2015). As shown in Table 3, total grade effect correlations with the prior year (lag 1) range from 0.598 (English) to 0.669 (science), persistent grade effects range from 0.444 (English) to 0.625 (Social Science), and mismatch from 0.478 (English) to 0.583 (science). Correlations decline slightly with each additional year as expected, suggesting modest changes in teacher effects and grading over time. Ninth-grade English teachers' grading effects are the least stable over time, while ninth-grade science and social science teachers' effects are the most stable. It is possible this reflects a greater emphasis on professional development in math and English, which are the subjects that receive more attention in state and district accountability policies.

Relationships of Teacher Grade Effects with Ninth-Grade Outcomes

The relationships of teacher grade effects with students' ninth-grade outcomes are shown in Table 4. We show estimates from a model with all four subjects together for simplicity; there are few differences by subject, with details provided in the online technical appendix. We normalized the teacher grade effects using the empirical standard deviations because the alternative standard deviations are only available for specific subjects, and we present pooled results for simplicity. As discussed in the section above, this is an underestimate of the true standard deviation of teacher grade effects; thus, the results given here underestimate the impact of having a teacher one standard deviation above the mean. A less conservative estimate can be obtained by multiplying the estimated effects by the ratio of the alternative standard deviations to the conservation standard deviations. This would increase the size of the teacher effects from those presented below by about 38 percent for total effects, 55 percent for persistent effects, and 65 percent for mismatch, if averaged across subjects.

Students who have ninth-grade teachers with total grade effects one standard deviation above the mean (as calculated when teaching other cohorts of students) end up with grades that are 0.279 GPA points higher in that teacher's class, on average, than students with more typical ninth-grade teachers. Persistent effects are based on students' *tenth-grade* course grades; even so, students who have ninth-grade teachers with high persistent effects also get higher average grades when they are in the ninth-grade class taught by the teacher, by 0.251 GPA points, compared to what they would get with an average ninth-grade teacher. In other words, students tend to get higher grades in both ninth and tenth grade with ninth-grade teachers with positive persistent effects, and lower grades with ninth-grade teachers with negative persistent effects.

Students also have higher ninth-grade course grades, on average, if they have a ninth-grade teacher that has positive mismatch, with grades in that teacher's class that are 0.290 GPA points higher for a teacher one standard deviation above the mean on mismatch, compared to a typical ninth-grade teacher. Perhaps surprisingly, they also have slightly higher grades in fall of tenth grade, by 0.013 GPA points. The mismatch comes from the difference between their spring and fall grades, indicating students of teachers with positive mismatch earn much higher grades than typical in the spring of ninth grade, so their tenth-grade grades did not match the high levels observed in ninth grade, even if they were slightly higher than expected. Note, the opposite is true for negative mismatch–students get much lower grades in ninth grade than their subsequent performance suggests were appropriate, and they end up with slightly lower tenth-grade grades than students who had ninth-grade teachers without grading mismatch.

Total and persistent teacher effects and mismatch are also associated with differences in broader ninth-grade outcomes. If the teacher with large positive total effects only influenced students' grades in their own class, we would expect that having a ninth-grade teacher one standard deviation above the mean in total effects would raise students' total core GPA by one-fourth of 0.279, which is 0.070. However, their core ninth-grade GPAs are higher by almost double that–by 0.127 GPA points. This suggests that having a ninth-grade teacher with higher total grade effects has positive spillover effects on grades in other classes.

Teachers' persistent effects show stronger relationships with broad ninth-grade outcomes than mismatch or total grade effects. Recall that ninth-grade teacher persistent effects are calculated based on their students' performance in tenth grade – the year after having the teacher. The fact that their students' grades are also higher/lower in other classes in ninth grade suggests that persistent effects capture teacher practices that raise or lower students' overall engagement in school in the year they have the teacher, as well as improving preparation for that subject the subsequent year. Students who have a ninth-grade teacher with persistent effects one standard deviation above the mean have ninth-grade GPAs that are 0.156 points higher than average. The difference between having a ninth-grade teacher with high and low persistent effects (a two standard deviation range) is over 0.3 GPA points. Ninth-grade failure rates, credits earned, and attendance are also better if students had teachers with positive persistent effects, and lower if they have teachers with negative persistent effects.

There is less spillover from having a ninth-grade teacher with more mismatch than total or persistent effects, but overall ninth-grade GPAs are higher by 0.113 points and attendance is higher by 0.2 percent if students have a ninth-grade teacher with strong positive mismatch (more lenient) than with negative mismatch (harsher), and they fail fewer courses overall.

Relationships of Teacher Grade Effects with Long Term Student Outcomes

Table 5 shows the differences in long-term outcomes for students who had one ninth-grade teacher one standard deviation above the mean in their total grade effects (higher ninth-grade grades), persistent grade effects (higher tenth-grade grades), and grading mismatch (more "lenient" grading). As discussed with Table 4, these estimates are based on the empirical standard deviations and likely under-estimate the effects of teachers one standard deviation above/below the mean on their students' long-term outcomes.

For all outcomes there are positive and significant coefficients associated with teachers' persistent effects. Having one ninth-grade teacher one standard deviation above the mean on persistent effects is associated with taking a slightly higher number of AP courses in eleventh and twelfth grade (an additional 0.129 courses), eleventh grade ACT scores that are 0.172 points higher, a graduating core GPA that is 0.140 points higher, and a 1.4 percentage point increase in the probability of graduating. The opposite is true for students with one ninth-grade teacher with negative grading effects. Thus, the difference in graduating core GPA for students who have a ninth-grade teacher one standard deviation above the mean in persistent effects compared to students who have a ninth-grade teacher one standard deviation below is 0.28 GPA points, with almost a three-percentage point difference in graduation rates. These estimates are conservative–teacher effects are likely 38 percent higher if we use the adjusted standard deviations, which would mean a difference in final GPAs of 0.39 GPA points from having a ninth-grade teacher with persistent effects a standard deviation above the mean versus one with persistent effects one standard deviation below the mean.

The long-term effects of having a ninth-grade teacher with grading mismatch are smaller than those of teachers' persistent effects. Having one ninth-grade teacher with positive mismatch one standard deviation above the mean is associated with a graduating core GPA that is 0.064 points higher than students who had average teachers, and an increase in the probability of graduating of 0.4 percentage points, holding constant teachers' persistent effects. The difference between having one ninth-grade teacher one standard deviation above the mean and a teacher one standard deviation below the mean on mismatch is a difference in graduating GPAs of 0.13 GPA points, and a difference in graduation rates of just under one percentage point (0.8 percent), holding constant teachers' persistent effects. If we use the less conservative estimates, teacher effects are 65 percent higher, which means a difference in graduating GPAs of 0.21 GPA points for students with ninth-grade teachers one standard deviation above the mean in terms of mismatch compared to students with teachers one standard deviation below the mean.

Robustness Checks

We conducted a number of analyses to investigate the potential for bias in our estimates resulting from tenth-grade teacher effects, school effects, and sorting. Details of these analyses are available in the supplementary online appendix.

Tenth-grade teacher effects: For conceptual and practical reasons described in the technical appendix, we did not control for potential tenth-grade teacher effects when estimating ninth-grade teacher persistent and mismatch effects. While these effects were estimated with an average of five tenth-grade teachers per ninth-grade teacher, for ten percent of ninth-grade teachers, all of their students enrolled with the same teacher in their subject in tenth grade. For these teachers, in particular, we were concerned that their persistent and mismatch effects were driven by the tenth-grade teacher as much as practices of the ninth-grade teacher. We hypothesized that if tenth-grade teachers were driving the estimates of mismatch and persistent effects for these teachers we would see: 1) higher variance in persistent and mismatch effects for these teachers, and 2) smaller correlations of total effects with persistent and mismatch effects (since total effects are estimated without consideration of tenth-grade grades). We did not find higher variance in persistent and mismatch effects for these teachers, and the correlations of their persistent and mismatch effects with their total effects were similar to those for other teachers. This gave us some confidence that tenth-grade teacher effects did not have a strong influence on our estimates for ninth-grade teachers.

School Effects: We did not include school fixed effects in the model for both theoretical and practical reasons. Theoretically, we thought that teachers who produced stronger outcomes in students were more likely to be in schools with stronger collaboration. Practically, many schools had only one or two ninth-grade teachers in a given subject making it difficult to disentangle school and teacher effects. The average school had 3.2 ninth-grade teachers for a given subject, the median number was just two, and a quarter of ninth-grade teachers were the only ninth-grade teacher of their subject in their school. We did not feel we could include school fixed effects and still get good estimates of ninth-grade teacher effects. However, we did want to estimate the degree to which school effects could be influencing our results. Therefore, we ran regression models with all subjects together (ensuring at least four teachers per school across the four subjects) and controlled for school and year fixed effects, taking the residuals as our new estimates of teacher effects. This process removes school effects, but it also removes the potential for spillover effects. It also likely over-controls for school effects, since many schools have small numbers of ninth-grade teachers. However, to the extent there are significant relationships with student outcomes it provides evidence that the teacher effects are not just a result of something happening at the school, but specific to different teachers within the same school.

Incorporating school and year fixed effects reduces the standard deviations of teacher effects, by 15 percent for total effects, 25 percent for persistent effects, and five percent for mismatch. This suggests that persistent effects are more strongly affected by school context than total effects or mismatch, perhaps because persistent effects are more strongly related to the conditions for teaching and learning in the school. Most of the relationships with short- and long-term outcomes remain significant, although the size of the coefficients are smaller, and mismatch shows slight negative relationships with tenth-grade course grades and ACT scores. Adjusting for school FEs gives us confidence that the persistent effects matter for students' long-term outcomes, while some of the mismatch effects may be due mainly to either spillover or overall school climate/structures.

Sorting: The validity of our estimated grade effects relies on the assumption that students are not sorted to teachers based on unobservable determinates of their grade outcomes. To attempt to assess this assumption, we follow the procedure outlined in Chetty (2014) to measure the degree of “forecast bias” introduced by sorting on one set of student characteristics available to us but not included in the main panel of controls: seventh-grade achievement variables, including GPAs and test scores.

We generate predicted versions of each of our three grade outcomes (ninth-grade spring grade, tenth-grade fall grade, grade difference) using the vector of seventh-grade variables, and examine the relationship of these predicted outcomes with each of our estimated grade effects. For all three of our estimated grade effects, we find these estimates to be extremely small – less than one hundredth of the size of the corresponding relationship between each grade effect and the observed grade outcome. This, combined with our robust baseline panel of controls, suggests that our estimates are not likely driven by sorting on unobservables. See the online appendix for further details on all of the supplementary analyses.

Discussion

Ninth-grade teachers have systematically different and quantifiable effects on students' course grades in their class, and in the same subject in the next grade. We find that our predicted effects, based on teachers' effects in years other than the year a student has the teacher, reliably correlate with students' actual gains in ninth- and tenth-grade course grades in the year left out of the estimate. Ninth-grade teachers with larger grade effects on some cohorts of students have similar effects on the grades of students in other cohorts, suggesting these are stable characteristics of teachers. Teachers with persistent effects on academic preparation, and those with idiosyncratic grading standards (large grading mismatch), not only influence students' grades in their own class and in the next year in the same subject, they also influence students' grades in other classes taken simultaneously, and students' long-term academic outcomes.

Students who have ninth-grade teachers with larger total grade effects, and especially teachers with larger persistent effects, have significantly stronger long-term outcomes, including eleventh grade ACT scores, graduating GPA, and high school graduation. The relationship of grading mismatch with later outcomes is more complicated and helps to explain the inconsistent findings of other researchers who found positive short-run effects on learning gains, but null or negative effects on long-term outcomes. Holding constant teachers' persistent effects, there are slight benefits to students' outcomes from more lenient grading (positive mismatch), which is consistent with the theory that grades can influence students' academic self-concept. But teachers who are lenient graders are less likely to have positive persistent effects than those who are harsh graders, and persistent effects are more strongly related to students' long-term outcomes than grading mismatch. This suggests larger benefits for students' long-term outcomes from teachers from whom it is harder to get high grades–but only if students' academic preparation actually increases. If they just get lower grades, there are negative long-term consequences overall. This has implications for teacher feedback on grading practices to maximize students' long-term outcomes.

The impact of ninth-grade teachers on students' long-term outcomes offers predictive validity of grading effects, and their size suggests the effects of one ninth-grade teacher on their students' outcomes can be substantial. Metrics around teachers' persistent effects on grades and their grading mismatch are an unused source of information that could be developed and used for feedback to improve teachers' effectiveness and their consistency. This could result in stronger short- and long-term outcomes for students, making it less likely a student would get a particularly strong or weak teacher by luck or by parental insight and influence. At the same time, efforts to improve consistency in grading should consider the scope of the problem and which teachers really need support.

Providing feedback to teachers on their impact on students' grades in the subsequent year could be valuable for instructional improvement. Metrics around teachers' persistent effects and mismatch could be used to prompt internal reflection, and collaborative discussions among teachers at the same grade level, and between teachers in the same subject at different grade levels. For example, low grade effects might lead a teacher or their supervisor to gather further data about how students experienced their class, relative to the classes of other teachers. Positive persistent grade effects could be used to inform peer mentoring, supported by data showing which teachers are especially effective for students' academic outcomes. Teachers' persistent effects could also be used in the evaluation of initiatives that aim to improve teaching effectiveness.

Teacher evaluation systems in many states use value-add metrics on assessments to gauge teacher effectiveness. Most teachers cannot receive value-add scores due to teaching in grade levels or subjects without test scores. Moreover, traditional value-add scores have been criticized for solely measuring tested skills, and leading teachers to narrow the curriculum to teach to the test (e.g., Jennings & Bearak, 2014; Koretz, 2017). Providing feedback on whether teachers have persistent effects on students' performance in subsequent classes could offer more comprehensive measures of teacher effectiveness for a much broader range of teachers.

Persistent effects incorporate all the different ways in which teachers prepare students for subsequent coursework–something not possible with assessments alone. They also allow for teachers to have effects in ways that may not be included on supervisors' rubrics, or that may be discounted by supervisors when evaluating teacher performance. At the same time, we do not advocate for using estimates of teacher grade effects for strict accountability measures or single indicators for personnel decisions. They could be one indicator, used in context, and considering the grading practices of the teachers students have in the subsequent year.

There are a number of implications for any efforts to improve grading practices:

First, any such efforts should focus on mismatch, rather than on teachers' average grades. Grade variations by teacher occur not only because some teachers are harsher or more lenient teachers than others, but also because some teachers are more, or less, effective at preparing students for subsequent academic coursework than their peers.

Second, in most schools, it may be more efficient to address grading discrepancies with the few teachers whose standards deviate considerably from others, than to spend valuable professional learning time working to improve the grading consistency of all teachers. Most teachers produce assessments of their students' skills that are within half of a GPA point of average. Despite grading mismatch, grades are not much more variable as measures of academic preparation than standardized tests**.** For instance, the SAT mathematics test is widely regarded as highly reliable, yet the measurement error for a given score (e.g., 31 points for an average mathematics score) is nearly as large as the average annual gain on their testing system (36 points in mathematics) (College Board, 2017; Kim, Moses & Zhang, 2018). For the SAT, 31 test score points is 0.30 standard deviations; thus, there is a 68% chance that a student's true score falls within 0.30 standard deviations (31 points) of the score they receive. In terms of grading mismatch, there is a 68% chance that a student will have a teacher who gives grades within 0.45 GPA points of the students' true skill levels–this is the empirical SD of mismatch multiplied by 1.65 to approximate the alternative standard deviation; 0.45 GPA points represents 0.38 standard deviations. Thus, teacher grading mismatch introduces just slightly more error into the assessment of students' academic preparation than the measurement error typically encountered in a highly reliable standardized test (0.38 versus 0.30), and less error than the test if we were to use the empirical standard deviation as a comparison.

Finally, these findings affirm the benefits of high expectations--teachers who positively impact subsequent academic performance tend to be harder graders. But they also show that giving low grades that fail to inspire greater effort, and better preparation does not help students in the long run. Thus, there is a balance—expectations should be challenging, but with enough support that students rise to the challenge and get good grades in the end.

Future research should explore whether the nature of tasks assigned, as well as the weighting given to these tasks, influences teachers' overall effectiveness. A concern that is not assessed in this study is whether the types of tasks teachers assign to students, and the way that they grade those tasks, impact student motivation and learning. There are debates about the degree to which teachers should incorporate behavior into grades, eliminate the use of zeros, etc. The current study solely compares teachers' effectiveness in improving students' grades the following year and whether these grades align with students' academic preparedness.

References

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665-676. https://doi.org/10.3102/0013189x09351981

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Prentice Hall.

Bastedo, M. N., Umbricht, M., Bausch, E., Byun, B.-K., & Bai, Y. (2023). Contextualized high school performance: Evidence to inform equitable holistic, test-optional, and test-free admissions policies. *AERA Open, 9*. https://doi.org/10.1177/23328584231197413

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review*, 22(4), 343-352. https://doi.org/10.1016/s0272-7757(02)00059-6

Bowers, A. J., Sprott, R., & Taff, S. A. (2012). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 77-100. https://doi.org/10.1353/hsj.2013.0000

Bransford, J.D., Brown, A., & Cocking, R. (2000). *How people learn: Mind, brain, experience, and school*. National Research Council. https://doi.org/10.17226/9853

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*(2), 123-142. https://doi.org/10.1111/j.1745-3984.1993.tb01070.x

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., ... & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848. https://doi.org/10.3102/0034654316672069

Caraway, K., Tucker, C. M., Reinke, W. M., & Hall, C. (2003). Self-efficacy, goal orientation, and fear of failure as predictors of school engagement in high school students. *Psychology in the Schools*, *40*(4), 417-427. https://doi.org/10.1002/pits.10092

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review, 104*(9), 56. https://doi.org/10.1257/aer.104.9.2593

Chicago Public Schools. (2017). *Professional grading standards and professional practices guidelines for Chicago Public School teachers.* Chicago Public Schools. https://www.cps.edu/globalassets/cps-pages/about-cps/policies/administrative-hearings/professional_grading_standards.pdf

College Board. (2017). C*ritical Evidence 2.1.1.a: SAT Suite Technical Manual Appendixes*. The College Board.

Cunha, F., & Heckman, J. J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources, 43*(4), 738-782. https://doi.org/10.1353/jhr.2008.0019

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance—a critical literature review*. Consortium on Chicago School Research.

Ferlazzo, L. (2011) Involvement or engagement?. *Educational Leadership, 68*(8), 10-14.

Fredricks, J.A., Blumenfeld, P.C., & Paris, A.H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109. https://doi.org/10.3102/00346543074001059

Geiser, S., & Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes.* Research & Occasional Paper Series: CSHE. 6.07. Center for Studies in Higher Education.

Gilraine, M., & Pope, N. G. (2021). *Making teaching last: Long-run value-added* (Working Paper No. w29555). National Bureau of Economic Research. https://doi.org/10.3386/w29555

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers?. *Educational Evaluation and Policy Analysis, 35*(2), 220-236. https://doi.org/10.3102/0162373712466938

Gwynne, J. A., & Moore, P. T. (2017). Chicago's Charter High Schools: Organizational Features, Enrollment, School Transfers, and Student Performance. Research Report. *University of Chicago Consortium on School Research*.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy, 126*(5), 2072-2107. https://doi.org/10.1086/699018

Jackson, C. K., Kiguel, S., Porter, S. C., & Easton, J. Q. (2024). Who benefits from attending effective high schools?. *Journal of Labor Economics, 42*(3), 717-751. https://doi.org/10.1086/724568

Jennings, J. L., & Bearak, J. M. (2014). “Teaching to the test” in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher, 43*(8), 381-389. https://doi.org/10.3102/0013189X14554449

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper No. w14607). National Bureau of Economic Research. https://doi.org/10.3386/w14607

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

Kelly, S. (2008). What types of students' effort are rewarded with high marks?. *Sociology of Education*, *81*(1), 32-52. https://doi.org/10.1177/003804070808100102

Kim, Y. K., Moses, T., & Zhang, X. (2018). Student-level growth estimates for the SAT suite of assessments. *College Board Statistical Report*. The College Board. https://satsuite.collegeboard.org/media/pdf/student-level-sat-suite-growth-estimates.pdf

Koedel, C., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47,* 180-195. https://doi.org/10.1016/j.econedurev.2015.01.006

Koretz, D. (2017). *The testing charade: Pretending to make schools better.* University of Chicago Press. https://doi.org/10.7208/chicago/9780226408859.001.0001

Lee, V. E., Smith, J. B., Perry, T. E., & Smylie, M. A. (1999). *Social support, academic press, and student achievement: A view from the middle grades in Chicago*. Consortium on Chicago School Research.

Mahatmya, D., Lohman, B.J., Matjasko, J.L., & Farb, A.F. (2018). Engagement across developmental periods. In S.L. Christenson et al. (eds.), *Handbook of Research on Student Engagement* (pp. 45-63). Springer. https://doi.org/10.1007/978-1-4614-2018-7_3

Mitchell, R. M., Kensler, L. A., & Tschannen-Moran, M. (2015). Examining the effects of instructional leadership on school academic press and student achievement. *Journal of School Leadership, 25*(2), 223-251. https://doi.org/10.1177/105268461502500202

Mozenter, Z. (2019). *Essays on the effects of teacher grading standards and other teaching practices* [Doctoral dissertation, University of North Carolina Chapel Hill]. https://doi.org/10.17615/ejqp-rn19

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher*, *42*(5), 259-265. https://doi.org/10.3102/0013189x13481382

Protik, A., & Walsh, E. (2013*). Does tracking of students bias value-added estimates for teachers?* (Working Paper No. 15) Mathematica Policy Research. https://www.doi.org/10.13140/RG.2.1.4005.9360

Stepner, M. (2013), VAM: Stata module to compute teacher value-added measures. http://fmwww.bc.edu/RePEc/bocode/v/vam.ado.

Online Technical Appendix for Variability in the Grades Teachers' Give: Teacher Grade Value-Add, Mismatch, and Long-term Effects

This appendix provides information on the following questions regarding the analyses presented in the main manuscript:

1) To what extent are tenth-grade teacher effects influencing estimates of ninth grade teacher effects?
2) To what extent are school effects influencing the estimates of teacher effects and their relationships with student outcomes?
3) What are the relationships of teacher grading effects with student outcomes for each of the subjects?
4) Could students be sorting into ninth-grade teachers based on unobservable characteristics that bias the estimates of ninth-grade teacher effects?

To what extent are tenth-grade teacher effects influencing estimates of ninth-grade teacher effects?

Our estimates of persistent and mismatch effects of ninth-grade teachers could be influenced by the effects of students' tenth-grade teachers. We considered including tenth-grade teacher fixed effects in our models but realized that our ninth-grade cohorts' tenth-grade classes often contained many students who were not part of the ninth-grade cohort and not part of our analyses. While these students could be included in the construction of variables representing classroom composition used as covariates, the estimates of tenth-grade teacher effects could only be based on students who had been in ninth grade the prior year. We also would have had to remove the 2014-15 cohort from our analysis, as well as the 2015-16 cohort, since the issue with incomplete teacher IDs would have been problematic for identifying the tenth-grade teachers of that cohort. Note that any tenth-grade teacher effects would be likely to make our identification of ninth-grade teachers less accurate, biasing the teacher effects downward, such that the ninth-grade teacher effects would be likely to show relationships with later outcomes.

The concern about tenth-grade teacher effects is larger the more that students from a particular ninth-grade teacher go on to all have the same teacher for tenth grade, or a small number of tenth-grade teachers. For ten percent of ninth-grade teachers, all of their students enrolled with the same teacher in their subject in tenth grade. For these teachers, we cannot discern whether the persistent and mismatch effects resulted from the practices of the ninth- or tenth-grade teacher. We hypothesized that if tenth-grade teachers were strongly driving the estimates of mismatch and persistent effects for these teachers we would see: 1) higher variance in persistent and mismatch effects, and 2) smaller correlations of total effects with persistent and mismatch effects.

We did not find higher variance in persistent and mismatch effects for ninth-grade teachers whose mismatch and persistent effects were based on a smaller number of tenth-grade teachers. Table A1 shows the standard deviation of ninth-grade teacher effects based on the number of tenth-grade teachers that their mismatch and persistent effects were based. There is not a consistent pattern.

We also expected that total effects would be less correlated with persistent and mismatch effects for ninth-grade teachers whose mismatch and persistent effects were based on only one tenth-grade teacher. The more that tenth-grade teacher effects drive the estimates of persistent effects and mismatch, the less they should be correlated with total effects which are calculated without consideration of students' performance in tenth grade. For persistent effects, the correlation was similar for teachers whose students went to only one ninth-grade teacher and those whose students went to multiple tenth-grade teachers (0.39 versus 0.36). For mismatch, the correlation was slightly smaller (0.57 versus 0.62). This provided some assurance that tenth-grade teachers were not strongly contributing to the estimates of ninth-grade teacher effects, particularly since it was most typical for ninth-grade teachers to send students to four or more tenth-grade teachers.

To what extent are school effects influencing the estimates of teacher effects and their relationships with student outcomes?

We did not include school fixed effects in the model for both theoretical and practical reasons. Theoretically, we thought that teachers who produced stronger outcomes in students were more likely to be in schools with stronger collaboration. Practically, many schools had only one or two ninth-grade teachers in a given subject making it difficult to disentangle school and teacher effects. We did not feel we could include school fixed effects and still get good estimates of ninth-grade teacher effects. However, we did want to estimate the degree to which school effects could be influencing our results.

To isolate the effects on grades that came from teachers net of any school effects we ran regression models with all subjects together (ensuring at least four teachers per school across the four subjects) and controlled for school and year fixed effects, taking the residuals as our new estimates of teacher effects. This process removes school effects, but it also removes the potential for spillover effects. It also likely over-controls for school effects, since many schools have small numbers of ninth-grade teachers. It also removes the potential effects of principals selecting strong ninth-grade teachers, and positive effects of ninth-grade team collaboration. It asks: What is the effect of a strong ninth-grade teacher, net of any schoolwide conditions that affect that teacher, and whose positive or negative effects do not spill over into their students' success in other classes? While that is a narrow question, the contrast is helpful for thinking about how much could be attributed to an individual teacher net of the broader school context, versus the effects of the school as a whole on the outcomes for students of individual teachers.

Incorporating school and year fixed effects reduces the standard deviations of teacher effects by 15 percent for total effects, 25 percent for persistent effects, and five percent for mismatch, see Table A2. This suggests that persistent effects are more strongly affected by school context than total effects or mismatch.

Most of the relationships with short- and long-term outcomes remain significant, although the size of the coefficients are much smaller. In addition, ACT scores are no longer significantly related to total or persistent effects, while the relationship of mismatch with ACT scores becomes slightly negative. The fact that most coefficients remain significant suggests that individual teachers have unique influences on students' long-term outcomes, not reflected in students' performance in their other ninth-grade classes, or the benefits of collaboration among teachers in the school. This gives us confidence that the persistent effects matter for students' long-term outcomes, while some of the mismatch effects may be due mainly to either spillover or overall school climate/structures. But the reduction in the size of the coefficients also suggests that a community of teachers may be particularly important, relative to just having strong individual teachers. It also suggests that spillover effects may also be an important contributor to long term outcomes; for example, encouraging or discouraging effort in other subjects, which then contributes to success in later years. The data do not allow us to parse how much is spillover and how much is a shared community or school culture that supports (or does not support) strong grades. Future research might try to better disentangle all of these sources of influence on the measurement of teacher effects and on student outcomes.

What are the relationships of teacher grading effects with student outcomes for each of the subjects?

The relationships of ninth-grade teacher grade effects with students' short- and long-term outcomes for each of the subjects are displayed in tables A4 and A5. The models were run in the same way described in the main manuscript for all subjects combined, but separately for each subject. As observed in the combined results in the main manuscript, the ninth- and tenth-grade outcomes are all significantly related to each of the teacher grade effects by subject, with the exception of students' tenth-grade course grades which are not significantly related to ninth-grade teacher mismatch in any subject or in the combined model. Almost all of the long-term outcomes that showed significant relationships with the combined ninth-grade teacher also show significant relationships with each of the subject-specific grade effects with two exceptions. The number of AP courses taken in eleventh and twelfth grade only shows significant relationships with persistent effects in math and social science, not English or science. ACT scores show significant relationships with ninth-grade teachers' total and persistent effects in all subjects except English.

Could students be sorting into ninth-grade teachers based on unobservable characteristics that bias the estimates of ninth-grade teacher effects?

The validity of our estimates of each teacher grade effect relies on the assumption that students are not sorted to teachers based on unobservable determinates of student grade outcomes. To attempt to assess this assumption, we can test the extent to which unobservable characteristics may be introducing bias into our estimates using a set of seventh-grade achievement variables not included in our primary control vector. These include student's seventh-grade GPA, seventh-grade percentile achievement in math and reading, and average math score of seventh graders in their school.

We estimate the "forecast bias" of this group of seventh-grade outcomes adapting the procedure introduced by Chetty et al. (2014). Chetty et al. (2014) documents mathematically how this approach can be used to measure the concept of forecast bias. Intuitively, we would not expect our estimated ninth-grade teacher value added measures to tell us anything about students' predicted grades (based on their seventh grade performance) unless there is unobservable sorting on the seventh-grade variables not already explained by our panel of controls.

Denoting our vector of seventh-grade outcomes as S_{it}^{*}, we construct residualized seventh-grade outcomes S_{it} following our standard procedure – regressing each of the seventh-grade outcomes on our baseline control vector X_{ijt}, including teacher fixed effects. We then use this residualized values to create predicted values of each of our three grade outcomes (ninth-grade spring grade, tenth-grade fall grade, difference between ninth- and tenth-grade grades): for example, constructing predicted values of students' ninth-grade spring course grades according to ${g^{S}}_{ijt} = \hat{\rho} S_{it}$. We fit separate models for each subject both when residualizing the seventh-grade outcomes (as per usual in our residualization procedure) and when generating the predicted grade outcomes.

We then estimate the extent of sorting based on seventh-grade outcomes by regressing each predicted grade outcome on the corresponding normalized teacher grade effect. In Table A6, we compare these estimates with estimates from the same specification using the actual observed grade outcomes. For each of the observed three student grade outcomes (students' ninth-grade spring grade, tenth-grade fall grade, and difference between the two grades), we see that a one standard deviation increase in each teacher grade-effect has a large and significant relationship, approximately equal in GPA points to the observed grade-effect in the main tables. On the other hand, the relationship of each teacher grade effect with the predicted grade outcomes based on seventh grade scores is very small – in all cases less than one hundredth of the size. In all cases, the magnitudes indicate that these effects are not meaningful and do not suggest a high degree of forecast bias in our estimates.

Table A1. Variance of Ninth-Grade Teacher Effects by Number of Tenth-Grade Teachers Receiving Students from the Teacher

Number of Tenth-Grade Teachers	Obs	Variable	N	Std Dev
1	609	Mismatch	609	0.246
		Persistent Effect	609	0.223
		Total Effect	609	0.269
2	591	Mismatch	591	0.277
		Persistent Effect	591	0.232
		Total Effect	591	0.305
3	668	Mismatch	668	0.265
		Persistent Effect	668	0.230
		Total Effect	668	0.293
4	740	Mismatch	740	0.273
		Persistent Effect	740	0.247
		Total Effect	740	0.316
5	749	Mismatch	749	0.253
		Persistent Effect	749	0.229
		Total Effect	749	0.287
6	696	Mismatch	696	0.233
		Persistent Effect	696	0.212
		Total Effect	696	0.265
7	495	Mismatch	495	0.252
		Persistent Effect	495	0.229
		Total Effect	495	0.254
8	346	Mismatch	346	0.282
		Persistent Effect	346	0.230
		Total Effect	346	0.292
9	197	Mismatch	197	0.285
		Persistent Effect	197	0.204
		Total Effect	197	0.319
10 or more	454	Mismatch	454	0.302
		Persistent Effect	454	0.197
		Total Effect	454	0.290

Table A2. Standard Deviation of Teacher Effects Controlling for School and Year Fixed Effects

Teacher Effect	Standard Deviation	Percent decline relative to empirical SD
Total	0.248	15%
Persistent	0.177	25%
Mismatch	0.256	5%

Table A3. Relationships of Teacher Effects with Outcomes Controlling for School and Year Fixed Effects

	Ninth-Grade Course Grade	**Tenth-Grade Course Grade**	**Ninth-Grade Core GPA**	**Ninth-Grade Fs in Core Courses**	**Ninth-Grade Attendance**	**Number of AP Courses**	**ACT Score**	**Graduating GPA**	**Diploma in 4 Years**
Total	**0.210**	0.004	**0.048**	**-0.046**	**0.055**	**0.039**	0.004	**0.017**	0.043
s.e.	*0.002*	*0.002*	*0.002*	*0.003*	*0.014*	*0.007*	*0.006*	*0.001*	*0.065*
pvalue	*0.000*	*0.093*	*0.000*	*0.000*	*0.000*	*0.000*	*0.522*	*0.000*	*0.506*
Persistent	**0.148**	**0.120**	**0.052**	**-0.052**	**0.044**	**0.088**	0.007	**0.032**	**0.356**
s.e.	*0.003*	*0.003*	*0.002*	*0.004*	*0.017*	*0.008*	*0.007*	*0.002*	*0.078*
pvalue	*0.000*	*0.000*	*0.000*	*0.000*	*0.009*	*0.000*	*0.280*	*0.000*	*0.000*
Mismatch	**0.244**	**-0.015**	**0.055**	**-0.059**	**0.045**	**0.055**	**-0.014**	**0.019**	0.032
s.e.	*0.002*	*0.003*	*0.002*	*0.004*	*0.016*	*0.008*	*0.006*	*0.002*	*0.074*
pvalue	*0.000*	*0.000*	*0.000*	*0.000*	*0.004*	*0.000*	*0.024*	*0.000*	*0.670*

Note: These teacher effects come from models that include all subjects together along with school and year fixed effects. As a result, spillover effects are controlled along with school effects.

Table A4. Ninth-Grade Teacher Effects on Ninth- and Tenth-Grade Outcomes by Subject

		English		Math		Science		Social Science	
		Coeff	std. err.	Coeff	std. err.	Coeff	std. err.	Coeff	std. err.
9th Grade Course Grade	Total Effect	0.286	0.010	0.272	0.010	0.268	0.009	0.304	0.012
	Persistent Effect	0.267	0.013	0.243	0.012	0.257	0.011	0.242	0.011
	Mismatch	0.318	0.012	0.283	0.010	0.285	0.010	0.285	0.012
10th Grade Course Grade	Total Effect	0.086	0.018	0.098	0.014	0.085	0.014	0.130	0.019
	Persistent Effect	0.252	0.016	0.226	0.011	0.235	0.012	0.246	0.011
	Mismatch	0.020	0.014	0.013	0.010	0.008	0.011	0.015	0.011
9th Grade Core GPA	Total Effect	0.137	0.009	0.123	0.009	0.112	0.008	0.149	0.012
	Persistent Effect	0.172	0.012	0.149	0.010	0.131	0.010	0.181	0.009
	Mismatch	0.143	0.009	0.105	0.008	0.101	0.008	0.114	0.010
Fs in Core Courses in 9th grade	Total Effect	-0.133	0.020	-0.103	0.015	-0.116	0.013	-0.149	0.020
	Persistent Effect	-0.188	0.023	-0.116	0.015	-0.090	0.016	-0.195	0.018
	Mismatch	-0.178	0.029	-0.096	0.018	-0.128	0.015	-0.124	0.018
9th grade attendance	Total Effect	0.428	0.078	0.152	0.076	0.319	0.069	0.227	0.081
	Persistent Effect	0.572	0.100	0.399	0.073	0.211	0.087	0.322	0.080
	Mismatch	0.421	0.102	-0.056	0.091	0.346	0.086	0.056	0.084
Credits earned in 9th grade	Total Effect	0.042	0.016	0.048	0.012	0.053	0.010	0.061	0.019
	Persistent Effect	0.075	0.025	0.047	0.014	0.056	0.012	0.093	0.021
	Mismatch	0.054	0.018	0.049	0.014	0.050	0.011	0.056	0.017

Note: This table corresponds with Table 4 in the main manuscript.

Table A5. Ninth-Grade Teacher Effects on Long-term Student Outcomes by Subject

		English		Math		Science		Social Science	
		Coeff	Std err	Coeff	Std err	Coeff	Std err	Coeff	Std err
N. of AP Courses in 11th & 12th grade	Total Effect	-0.042	0.094	0.035	0.060	-0.005	0.065	0.229	0.066
	Persistent Effect	0.057	0.094	0.195	0.072	-0.041	0.068	0.304	0.072
	Mismatch	0.056	0.152	-0.062	0.062	0.055	0.091	0.157	0.069
ACT Score	Total Effect	0.082	0.043	0.159	0.027	0.099	0.029	0.087	0.037
	Persistent Effect	-0.005	0.062	0.198	0.036	0.283	0.035	0.114	0.042
	Mismatch	-0.005	0.055	0.116	0.035	0.005	0.031	-0.026	0.038
Graduating Core GPA	Total Effect	0.104	0.012	0.090	0.009	0.077	0.007	0.112	0.014
	Persistent Effect	0.179	0.014	0.137	0.009	0.109	0.009	0.155	0.010
	Mismatch	0.081	0.009	0.059	0.007	0.057	0.007	0.064	0.008
Prob. of Any Diploma in 4 Years	Total Effect	0.50	0.24	0.43	0.19	1.00	0.19	1.23	0.30
	Persistent Effect	2.16	0.29	1.12	0.23	1.41	0.23	1.38	0.28
	Mismatch	0.08	0.29	-0.07	0.23	0.76	0.25	0.76	0.29

Note: This table corresponds with Table 5 in the main manuscript.

Table A6. Estimates of Bias Due to Sorting based on Seventh-Grade Achievement Variables

Relationships of teacher grade effects with actual and predicted grades (in GPA points)

Teacher Grade Effect	Ninth-Grade Spring Grade (1)	Predicted Ninth-Grade Spring Grade Based on Seventh Grade (2)	Tenth-Grade Fall Grade (3)	Predicted Tenth-Grade Fall Grade Based on Seventh Grade (4)	Grade Difference (5)	Predicted Grade Difference Based on Seventh Grade (6)
Total (s.e.)	0.2786 (.0019)	0.0002 (.0002)				
Persistent (s.e.)			.2319 (.0020)	.0007 (.0002)		
Mismatch (s.e.)					.2745 (.0021)	.0019 (.00004)

Note: This table displays estimates from regressing both the observed grade outcomes and the grade outcomes predicted using the vector of seventh-grade variables on each estimated grade effects. The grade effects are normalized, so that the coefficient displayed corresponds to a one standard deviation increase in estimated grade-effect. Each model is estimated separately and includes subject fixed effects. The predicted values are generated using residualized versions of a vector of seventh-grade variables – more information on this procedure can be found above.