NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of North Carolina at Chapel Hill, University of Texas at Dallas, and University of Washington

How Much Do Early Teachers Matter?

Dan Goldhaber
Zeyu Jin
Richard Startz

WORKING PAPER No. 264-0422 • April 2022

How Much Do Early Teachers Matter?

Dan Goldhaber
American Institutes for Research/CALDER
University of Washington/CEDR

Zeyu Jin
American Institutes for Research/CALDER

Richard Startz
University of California, Santa Barbara

Contents

Acknowledgments

This research was supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders. Any errors are attributable to the authors.

We are grateful to Aurora Delaigle for code used in density estimation and to James Cowan, Eric Hanushek, Shelly Lundberg, and Jesse Rothstein for comments on an earlier draft of this paper.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

How Much Do Early Teachers Matter?

Dan Goldhaber, Zeyu Jin, Richard Startz

CALDER Working Paper No. 264-0422
April 2022

Abstract

We present new estimates of the importance of teachers in early grades for later grade outcomes, but unlike the existing literature that examines teacher "fade-out," we directly compare the contribution of early-grade teachers to later year outcomes against the contributions of later year teachers to the same later year outcomes. Where the prior literature finds that much of the contribution of early teachers fades away, we find that the contributions of early-year teachers remain important in later grades. The difference in contributions to eighth-grade outcomes between an effective and ineffective fourth-grade teacher is about half the difference among eighth-grade teachers. The effect on eighth-grade outcomes of replacing a fourth-grade teacher who is below the 5th percentile with a median teacher is about half the underrepresented minority (URM)/non-URM achievement gap. Our results reinforce earlier conclusions in the literature that teachers in all grades are important for student achievement.

1. Introduction

Students typically spend 13 years in K–12 education. Although there are important outcomes realized in early years, society is also concerned with longer term academic results. Teachers are generally considered to be the most important schooling resource, based on an abundance of evidence on their effects on both short-run test (e.g., Aronson et al., 2007; Goldhaber et al., 1999; Rivkin et al., 2005) and non-test (Jackson, 2018; Kraft, 2019) outcomes as well as longer term outcomes, such as high school graduation, college going, and labor market participation and earnings (Chetty et al., 2014).[1] Indeed, teachers are so important that, according to one estimate, a child in poverty who has a good teacher for 5 years in a row would have learning gains large enough, on average, to close completely the achievement gap with higher income students (Hanushek et al., 2005). At the same time, there is abundant evidence that much of the teacher contribution to student test-score gains largely fades out within just a grade or two (e.g., Jacob et al., 2010; Kane & Staiger, 2008; Kinsler, 2012; Konstantopoulos & Chung, 2011; Master et al., 2017; Rothstein, 2010). How can these seemingly divergent findings be reconciled?

We argue that the reason existing literature finds a very large degree of fade-out is due to the choice of benchmark. Essentially, the finding is that even when a teacher's contribution to student tests is large in their own grade, the majority of that contribution appears to be gone by the time students take tests in later grades. Thus, measures of fade-out in the existing literature compare the performance of an early-year teacher's students on tests given in the teacher's own grade to the performance of those students on a test given in a later year that is still detectible in terms of student test scores. This drop in apparent teacher contributions measured at late versus early years has been labeled "fade-out." The inference is that much of what the "unusually good" early-year teacher imparted has been lost by later years. The conclusion is that even excellent early-year teachers are not all that important if we care about long-run outcomes. This conclusion has implications for teacher accountability, pay, and how we think about investments in early-grade teachers (Leonhardt, 2010).

But the fade-out findings suggesting that investment in teachers may not have much of a long-run return also contradict the evidence that looks directly at long-term outcomes (e.g., Chetty et al., 2014). To the extent that we care about outcomes at or close to when students leave school, the fade-out findings might suggest that the importance of the teacher contribution had been much exaggerated, at least for early-year teachers. The fade-out literature has, to oversimplify, estimated value-added for teachers using both late-grade and early-grade tests and then regressed the former on the latter, using careful adjustments for a variety of econometric

[1] In turn, these teacher-quality effects are estimated to have consequential impacts on individual and aggregate economic well-being. Chetty et al. (2014), for instance, estimate that having a teacher with a 1 *SD* higher value-added adds about $230,000 in the present value of lifetime earnings to the students in that teacher's class.

issues. A string of influential papers (described in Section 2) finds that much of the contribution of early-grade teachers to test achievement is no longer detectible in terms of later grade student test scores. But a key issue with this approach to measuring the fade-out of teacher effects is that a teacher in Grade *t* could have large effects on student achievement in, for example, Grade *t+2* relative to Grade *t+2* teachers even if it is a small effect relative to the impact that the teacher has on Grade *t* student achievement.

In our view, early-year test scores are the wrong benchmark for consideration of long-run results. We take a student's performance *on a late-year test* and divide it up among the teachers the student had in that late year and the student's early-year teacher and all the teachers in between. We then compare the contributions of early-year teachers to the contributions of late-year teachers on the late-year outcomes. While the contributions of early-year teachers on long-run outcomes are smaller than those of late-year teachers, they remain substantial. In other words, the findings in the literature that early-grade teachers are important is supported by our empirical estimates.

The key to understanding why we get different results is the realization that the skills tested at various grade levels changes across grades and that not all material that we care about is tested. Suppose some fourth-grade teachers are effective at helping students memorize multiplication tables while, in contrast, other fourth-grade teachers foster in their students the ability to figure out more abstract mathematical concepts. Early-year tests check the ability to multiply but don't test "loving math." Eighth-grade tests, in turn, don't test the times tables but do test algebra—learning algebra being aided by an understanding of more abstract concepts. The students of the first teacher score well on the early test but not the later test, and vice versa for the students of the second teacher. In contrast, we compare eighth-grade test results for the students of fourth-grade teachers to those of the students of eighth-grade teachers, which is an apples-to-apples comparison.

The standard way to ask whether a good teacher contributes much more than a weak teacher is to look at the distribution of teacher effects and ask whether the dispersion is "large." Specifically, we look at the distribution of eighth-grade teacher effects (on eighth-grade tests) and the distribution of fourth-grade teacher effects (again on eighth-grade tests). The first measures we compare are the standard deviations of the two distributions. Not only is the fourth-grade standard deviation a substantial fraction of the eighth-grade standard deviation, but the fourth-grade standard deviation is also large compared to the underrepresented minority/nonminority (URM/non-URM) gap.

We also present other measures, such as the "Hanushek question" of what happens when a teacher in the tail of the distribution is replaced with an average teacher.[2] That the tails of the distributions in **Figure 1** are leptokurtic (as we describe later on) is quite important for this type of policy, because the heavy tails mean that the estimates of the effects of policies aimed at the tails of the teacher distribution are greater than estimates that assume that effectiveness is normally distributed. We use a kernel-smoothing algorithm developed by Delaigle and Meister (2008) and Delaigle et al. (2008) that allows us to make nonparametric estimates of the ability distribution appropriately adjusted for measurement error. As an example, using our estimates of teacher effectiveness while assuming normality, retaining a 95th-percentile fourth-grade teacher has about one third of the effect on eighth-grade outcomes as does retaining a 95th-percentile eighth-grade teacher. However, using our nonparametric estimates, the relative importance rises from one third to 70%.

We find results quite similar to those found by earlier researchers when we use our data and their methods, but less fade-out of teacher effects on later grade test scores. In particular, while studies that focus specifically on teacher fade-out find that only about a quarter of tested learning persists into the following year (e.g., Jacob et al., 2010) and 14% to19% persists two grades later (Kinsler, 2012; Jacob et al., 2010), we find that the variation in eighth-grade outcomes contributed by fourth-grade teachers is about 45% of that contributed by eighth-grade teachers. This conclusion is readily apparent in Figure 1, which shows the distribution of fourth- to eighth-grade teacher contributions to end-of-year student eighth-grade test scores. The dispersion of the eighth-grade teachers is greater than that of earlier grade teachers, but it is also true that the earlier grade teacher distributions represent a large fraction of the eighth-grade teacher variation.

Before turning to our model, we provide a brief review of the literature, with an emphasis on how our estimates are related to what has been done in the past. We then describe the data and explain our methodology, followed by a more detailed discussion of the results and a conclusion.

2. Models and Related Literature

A standard finding in the literature is that teachers are the most important schooling variable influencing short-run (i.e., year to year) gains in students' test achievement (e.g., Aaronson et al., 2007; Goldhaber et al., 1999; Rivkin et al., 2005). There is also evidence that teachers influence longer run outcomes, such as postsecondary schooling and labor market earnings (Chamberlain, 2013; Chetty et al., 2014). Thus, it is somewhat puzzling that researchers exploring the extent to which teacher impacts on test scores in one grade persist into later grades

[2] Hanushek (2011), for instance, has estimated that replacing the bottom 5%–8% of teachers with average teachers could increase the aggregate GDP of the United States by a present value of $100 trillion.

find that a significant portion of teacher effects on student tests in early grades cannot be detected in later grades—that is, that they appear to "fade out."[3]

A number of studies focus on assessing the contribution of teachers in Grade *t* compared with later grades (*t+1*, *t+2*, etc.) by including lagged value-added estimates or teacher indicators directly in student achievement models and assuming a constant rate of decay of teacher effects (e.g., Kane & Staiger, 2008; Konstantopoulos & Chung, 2011; Lockwood et al., 2007; Rothstein, 2010). These studies find year-to-year persistence in the range of 0.2 to 0.5. Importantly, however, the approach to measuring persistence in of all of the studies is to compare the contribution that teachers make to student achievement in Grade *t* relative to their contribution to student achievement in some later grade.[4]

We begin by providing some intuition (and formalize it further on) about challenges inherent in measuring how much knowledge students have as they advance from grade to grade (Rose, 2021). **Figure 2** illustrates that student learning in one grade builds on a foundation of the concepts taught in prior grades but also that what is taught in one grade may not be well measured by end-of-grade tests (Ballou, 2009; Koretz, 2009). In the figure, note that the size of the box of material covered by tests (the area inside of the dashed boxes) is, by assumption, consistent across grades. But the amount of knowledge expands from grade to grade, so the share of total knowledge that is tested declines as students advance from one grade to the next. Note also that some knowledge taught in early grades may be covered by early-grade tests but not later grade tests.[5] Finally, the figure depicts the fact that the contributions to knowledge or skills made by teachers in early grades may not be tested in those grades, despite being important for learning that occurs in later grades and is tested.[6] Consequently, comparing the effects of teachers based on their contributions to learning assessed by Grade t tests and comparing that to the learning associated with early-grade teachers that is identified by later grade tests may miss contributions. In particular, given that any domain of knowledge expands as students progress from grade to grade and the fact that tests (indicated by the white dashed boxes) can only sample from this domain of knowledge, later grade tests are likely to be a relatively smaller sample of accumulated knowledge and thus will less accurately assess the full breadth of students'

[3] As an example, if only half the contribution of a teacher persists into the following grade and if later fade-out is geometric, then even the nation's best teacher will have contributed virtually nothing to outcomes by their students' high school graduation.

[4] For a related study of persistence of teacher effects at the college level, see Carrell and West (2010).

[5] In Figure 2, the Grade *t+2* tests only cover a small portion of the material from Grade *t* (the portion inside the green-shaded dashed box in Grade *t+2*).

[6] In Figure 2, the solid boxes in Grades *t+1* and *t+2* that are in the green-shaded areas represent material not covered by the Grade *t* test, which is covered by later grade tests. To be more concrete, the contribution of teachers in fourth grade who provide students with a solid foundation in understanding factors in the fourth grade will show up on tests that cover fractions in the sixth grade. But, on the other hand, identifying the names of particular shapes (e.g., a triangle), another typical fourth-grade skill, may not show up on sixth-grade tests; consequently, the contribution of a teacher who helps students understand shape identification may appear on fourth-grade tests but not sixth-grade tests. For more information on skills that may be taught in different grades, see http://www.corestandards.org/Math/.

knowledge. And, as Cascio and Staiger (2012) point out, observed teacher fade-out (or fade-out of any educational intervention) could be an artificial consequence of how tests are scaled.[7]

We describe this aspect more formally in Equation 1, borrowing elements from a framework described in Jacob et al. (2010).[8] Their basic model is

$$\tau_{t'} = \lambda\tau_t + \varepsilon_t \quad (1)$$

where $\tau_{t'}$ is the average Year t' value-added score of students who had a particular teacher in Year t and τ_t is that value-added associated with that teacher. Jacob et al. (2010) use an instrumental variable approach that corrects for τ_t being a generated regressor that contains measurement error. The authors also provide an estimate of 2-years-apart persistence that shows further fade-out, but at a rate slower than geometric.

We extend the theoretical framework presented in Jacob et al. to illustrate two issues: (a) that our measure of fade-out is conceptually different than what has been previously estimated and (b) that there is a particular reason why estimates that focus on the dynamics of measured value-added may overstate fade-out. Since the objective is to provide intuition, we assume here that there are only two grades—fourth and eighth, although the empirical model includes the intervening grades—and a very limited set of skills.

Fourth-grade teachers impart three skills—memorizing multiplication tables (subscript M), loving math (subscript L), and good study skills (subscript S)—in amounts τ_{M_4}, τ_{L_4}, and τ_{S_4} as shown in Equation 2. Eighth-grade teachers impart skills in algebra (subscript A), loving math, and good study skills but do not teach multiplication.

All components last forever, and all components are mutually independent (to simplify the exposition) within and across grades. Let us assume that (a) in the fourth-grade loving math does not contribute to memorizing multiplication tables but does contribute in the long run to learning algebra (an eighth-grade skill) and that (b) fourth-grade tests examine only multiplication whereas eighth-grade tests examine only algebra,[9] but study skills help students in both grades.

[7] Cascio and Staiger explore this issue of the test domain-related fade-out (i.e., whether fade-out *appears* because of changes in the variance of test scores as students advance through school). They do find some evidence of this source of fade-out but conclude that the increased test-score variance explains relatively little of the empirically observed diminishment of early schooling interventions, such as having a very effective or ineffective teacher. As we describe further on in this article, because our preferred measure of fade-out is only based on late-grade tests (where traditional measures compare tests instruments across grades), our measure should not be at risk from this type of scaling issue.

[8] Jacob et al. (2010) describe how the persistence of long-term teacher effects can be recovered through instrumental variable estimation of student achievement and estimate that only about 0.20–0.25 of teacher effects persist after 1 year.

[9] This is, thus far, conceptually equivalent to the Jacob et al. (2010) "short-term" and "long-term" knowledge.

Ignoring other factors that will influence student achievement, a test score at the end of fourth grade will be

$$y_4 = \tau_{M_4} + \tau_{S_4}$$

and a test score at the end of eighth grade will be

$$y_8 = \tau_{A_8} + \tau_{L_8} + \tau_{S_8} + \tau_{L_4} + \tau_{S_4}$$

If we regress the latter on the former, we get

$$\frac{\sigma^2_{S_4}}{\sigma^2_{M_4} + \sigma^2_{S_4}} \tag{2}$$

Even though love of learning imparted in fourth grade persists, it does not show up in the numerator because this skill is not picked up on the fourth-grade test and therefore does not contribute to *measured* fourth-grade value-added. And the converse is also possible: Knowledge that teachers contribute to students in the fourth grade that is covered by fourth-grade tests may not be covered by eighth-grade tests (e.g., multiplication tables) directly or indirectly through an assessment of student skills that build upon the earlier foundation of knowledge.[10] In light of this possibility, we note that to the extent that there are skills taught in fourth grade that are not tested in eighth grade, our estimates will be too conservative in estimating the importance of early-year teachers.

In our approach, we allocate eighth-grade outcomes among the contributing teachers. Eighth-grade teachers are credited with $\tau_{A_8} + \tau_{L_8} + \tau_{S_8}$, and fourth-grade teachers are credited with $\tau_{L_4} + \tau_{S_4}$. Our measure compares the relative contributions (Exhibit 3).[11]

$$\frac{\sigma^2_{L_4} + \sigma^2_{S_4}}{\sigma^2_{A_8} + \sigma^2_{L_8} + \sigma^2_{S_8}} \tag{3}$$

If one is concerned about the contribution of early-year teachers to later learning, Equation 3 provides the correct conceptual measure. But note also that the fraction in Equation 3 may be either larger or smaller than the fraction in Equation 2. Thus, it may be that one reason for our finding of persistence greater than has been previously reported is that early-year teachers impart important skills that are only measured in future years. Indeed, as we show in Appendix

[10] Note that our Equation 2 is conceptually similar to Jacob and colleagues' Equation 12 (2010; p. 921) in that it is comparing the ratio of the contributions that fourth-grade teachers make toward material tested in the eighth grade (the numerator) to the combination of the contributions that fourth-grade teachers make to fourth- and eighth-grade tested material (the denominator).

[11] And note that because we are only using eighth-grade test outcomes, we sidestep the test scaling issue identified in Cascio and Staiger (2012).

A, when we employ the Jacob et al. (2010) approach to measuring teacher fade-out, we find results that closely parallel theirs.

Earlier work derives teacher effects by regressing student test scores in Grade *t* on teacher indicator variables and controls, including student test scores in Grade *t–1*. A notable exception is Kinsler (2012), who uses an approach more closely related to what we do. Illustrating Kinsler's method with only two grades (Equation 4),

$$\begin{aligned} y_8 &= \sum_{j^8 \in J^8} \tau_{j^8}^8 I_{j^8} + \lambda \sum_{j^4 \in J^4} \tau_{j^4}^4 I_{j^4} \\ y_4 &= \sum_{j^4 \in J^4} \tau_{j^4}^4 I_{j^4} \end{aligned} \tag{4}$$

One can think of Equation 4 as using the second part to identify early-year teacher value-added, which is then "fed into" the first part, identifying λ. The value λ, which provides the traditional measure of how early-year knowledge remains in later years, is derived from an estimate, with fourth-grade value-added on the right, so the same issues as in Equation (2) exists.[12]

Our specification is essentially the first part of Equation 4 (see Equation 5) but without forcing fourth-grade teacher-effect estimates to reflect early as well as later outcomes and without the λ parameter that assumes constant decay:

$$y_8 = \sum_{j^8 \in J^8} \tau_{j^8}^8 I_{j^8} + \sum_{j^4 \in J^4} \tau_{j^4}^4 I_{j^4} \tag{5}$$

We replace estimation of fourth-grade teacher effects on fourth-grade scores with the estimation of fourth-grade teacher effects on eighth-grade scores and then compare the distributions of τ^8 and τ^4, thus avoiding the issues described earlier. If the dispersion of teacher fixed effects for early-year teachers is small (compared to the dispersion for current-year teachers), then it makes relatively little difference whether a student has a good or bad teacher early on—the reverse being true if the dispersion among early-year teachers is relatively large. We would interpret the former case as one of substantial "fade-out."

$$y_8 = \sum_{j^8 \in J^8} \tau_{j^8}^8 I_{j^8} + \sum_{g \in G} \sum_{j^g \in J^g} \tau_{j^g}^g I_{j^g} + X\beta \tag{6}$$

[12] But continuing along with our simplified example, the estimates of fourth-grade value-added will be influenced by τ_{M_4} in the upper equation but not in the lower one. So, some element of $\sigma_{M_4}^2$ will be picked up.

Equation (6) is extended from the previous specification. We measure previous teacher-grade effects on eighth-grade student scores by including all of the earlier teacher dummies, I_{jg}. For each Grade 4–8, we acquire a set of teacher effects, τ^g. Importantly, X_{ijt} includes third-grade math test scores as well as student characteristics (gender, learning disability status, gifted, free or reduced-priced lunch [FRL] eligibility, URM status, Asian-Pacific Islander indicator, and participation in special education and English learner programs). The standard errors for each teacher-grade effectiveness are obtained by bootstrapping.

Equation (6) implicitly assumes that the impact of teachers on student achievement in one grade does not affect the assignment of students to subsequent teachers. This may be implausible. In the Section 5 (on robustness), we present Monte Carlo evidence that suggests that our estimates are conservative.

3. Data, Sample, and Measurement Issues

For our analyses, we use administrative data on individual public school students and teachers in Grades 4–8 provided by the Washington State Office of Superintendent of Public Instruction. This data include nine cohorts of students who began in fourth grade in the 2006–07 through 2014–15 school years, who could be followed to their eighth-grade year, and for whom there was information about their mathematics test achievement in third grade.[13] The data also contain detailed information on individual student background variables, including gender; race/ethnicity; learning disability status; FRL eligibility; and participation in gifted/highly capable, limited English proficiency, and special education programs. These student-level variables (for students when they are in the third grade) are used as control variables in the value-added models described earlier.

We include students who match with exactly one math teacher in each grade and school year and eliminate students with missing background information.[14] Between 2006–07 and 2008–09, we link students in Grades 4–6 in elementary schools to their classroom teachers through a proctor field in the state assessment file, but in the 2009–10 school year, students can

[13] Students were tested based on either the Washington Standards of Student Learning (WASL), the Measures of Academic Progress (MAP), or the Smarter Balanced Assessment (SBA), depending on the year. The WASL was the state assessment used in the 2005–06 to 2008–09 school years, and the MAP was used in 2009–10 to 2013–14, when it was replaced by the SBAC test, which is still used. The SBA test was designed to be computer adaptive, but the WASL and MAP tests were not. About one third of schools in the state participated in a pilot of the SBA in 2013–14, and the state did not collect test scores from students in these schools for this school year. Thus, current test scores are missing for students in these schools in 2013–14, and prior test scores are missing for students in these schools in 2014–15.

[14] As an alternative approach to eliminating students with missing information, we impute values for each variable and create missing indicators. The findings on teacher effectiveness that we report further on in this article are quite similar to those with missing values imputed. We are happy to provide these results upon request.

be linked to their teachers using a unique classroom ID in the state's Comprehensive Education Data and Research System (CEDARS) database.[15]

Because we are regressing eighth-grade student achievement on multiple grades of teacher indicators, it is important to identify whether these teachers are networked together in the sense that they share common support; if they do not, then the teacher effects are not fully identifiable (that is, one cannot distinguish the effects in one group of networked teachers from the effects of teachers in a different independent network; Reardon & Raudenbush, 2009). Moreover, sparsely connected networks can lead to biased teacher-effect estimates (Jochmans & Weidner 2019).[16] To satisfy the necessity and sufficiency of identification conditions (Abowd et al., 2002) on estimating multiple high-dimensional fixed effects, we sample the largest group of teachers who are connected by students.[17] We then estimate teacher-grade effects based on the largest connected estimable group.[18] **Figure 3** demonstrates some intuition about connected groups. The left panel of the figure shows the teachers that students are assigned to, represented by letters, as they progress from grade to grade (each column is a grade). The right portion of the figure shows how the progression of students to different teachers creates three distinct connected groups. Specifically, each vertex represents a single teacher in a grade, and each edge represents the connection of teachers through the students who they have in their classrooms (i.e., two teachers are connected if and only if students from the early-grade teacher feed into the adjacent later teacher's classroom). Therefore, we obtain the groups based on the teacher connectivity and student mobility in Grades 4–8.

After we restrict the data based on connectedness, about 61% of students and 85% of teachers remain within the sample. Specifically, we have a sample of 173,858 unique students and 620,499 teacher-year observations (15,981 unique teachers). **Table 1** provides descriptive statistics for students. Column 1 are for the largest connected group of teachers (and for whom we have third-grade math scores (which are standardized across all third graders in any connected group). Column 2 includes students who are not in the largest connected group of teachers. And Column 3 is the difference across these two mutually exclusive categories.

---

[15] Note that the "proctor" variable was not intended to be a link between students and their classroom teachers, so this link may not accurately identify those classroom teachers, and while the Comprehensive Education Data and Research System (CEDARS) data include fields designed to link students to their individual teachers, based on class schedules, the limitations of reporting standards and practices across the state may result in ambiguities or inaccuracies around these links.

[16] This issue also arises in other educational contexts, such as when trying to estimate the effects of school principals (Bartanen & Husain, 2021) and teacher education programs (Mihaly et al., 2013).

[17] For the algorithm for determining a connected subset, we refer to the note of Weeks and Williams (1964). This algorithm finds connected components in which all differences between teacher-grade effects are estimable. The largest connected group is recommended and usually is sufficiently large for proper analysis.

[18] We use a computationally efficient method introduced by Gaure (2013) to estimate high-dimensional fixed effects (i.e., teacher-grade effects) using ordinary least squares after the largest connected estimable group has been selected. Note that the largest estimable group contain sufficiently large samples to estimate teacher-grade effects. The standard errors of fixed effects are obtained through bootstrap.

Students in the largest connected group are far more likely to be advantaged according to their third-grade test scores or FRL status,[19] and they are far less likely to be students of color. This is not surprising given that students who are more mobile tend to be more disadvantaged (Goldhaber et al., 2021), have poorer academic outcomes, and are less likely to be in the sample given that they have to have both third- and eighth-grade tests to be included.[20]

We use the connected sample just described to estimate the teacher value-added models (according to Equation (6)). The resulting teacher-grade value-added estimates, $\hat{\tau}_{jg}$, are noisy in that they include sampling error (Jacob & Lefgren, 2008; Kane & Staiger, 2008). We first follow standard practice and make an empirical Bayes correction to the estimated distribution of teacher value-added estimates. To remove the sampling error from the variance of estimated teacher-grade effects, we first estimate the values of $\hat{\sigma}_{jg}$, the standard errors on $\hat{\tau}_{jg}$, using the parametric bootstrap.[21] Then we remove the average sampling error, $\frac{1}{n}\sum \hat{\sigma}_{jg}{}^2$, from the variance of $\hat{\tau}_{jg}$, where n is the number of teachers in each Grade g. Finally, we get the corrected spread Σ of the density of teacher-grade effects from the following (Exhibit 7):

$$\Sigma = \sqrt{var\left(\hat{\tau}_{jg}\right) - \frac{1}{n}\sum \hat{\sigma}_{jg}{}^2} \tag{7}$$

Importantly, the standard approach is to adjust the teacher effectiveness distribution, assuming that it is Gaussian. Goldhaber and Startz (2017) suggest that for a number of data sets, departures from normality are not substantively important. But Gilraine et al. (2021) reach a different conclusion, finding that some value-added estimates from some data sets are approximately Gaussian whereas other data sets give substantively non-Gaussian results.

As we show in the following section, our estimates of the distribution of teacher value-added are highly non-Gaussian.[22] Thus, we also employ a nonparametric estimator (Delaigle et al., 2008; Delaigle & Meister, 2008), which provides a kernel-smoothing estimate that corrects

[19] Note that we standard normalize student test scores within grade and year so that the difference between Columns 1 and 2 in tests represents the average placement in the third-grade math distribution for students who are in the connected group and nonconnected groups.

[20] While we do not report it, the teachers in the connected sample also differ from those who instruct students in math but are not in the connected sample. In particular, they have somewhat higher levels of teaching experience. Again, this is not surprising given that teachers are more likely to be connected with each other through students when they spend more time in the state's teacher labor market.

[21] We employ a parametric bootstrap to keep the connectedness in each iteration. We also perform a clustered bootstrap within the eighth-grade classroom. The clustered bootstrap has smaller average standard errors in each grade than does the nonclustered bootstrap, which implies that the empirical Bayes corrected teacher value-added distributions we utilize are more conservative than using clustering.

[22] The p values for the Jarque-Bera test for normality are 0 to all reported decimal places. While the distributions are roughly bell-shaped, the deviation from normality is large enough to matter for policy questions such as replacing teachers in the lower tail. When we account for the role of measurement error in estimated teacher effects (see results in Section 4), we will see that the deviation from normality persists.

for measurement error. We leave the details to the original article, but effectively, Delaigle's method is a kernel smoother that makes an adjustment for the sampling error. The result of Delaigle's method is a density estimate on a set of grid points; four of the distributions are shown in Figure 1. We can then estimate the standard deviation or quantiles of the distribution numerically.

4. Results

Before delving into our primary findings on fade-out, it is worth noting that when we utilize the Jacob et al. (2010) methodology of estimating teacher persistence (see Appendix A), we obtain estimates of teacher fade-out that are comparable to the original Jacob et al. estimates and far smaller than those we cover in the discussion that follows. In addition, it is useful to have some context for how to think about the substantive size of the following estimates of the importance of variation in teacher effects. One such standard of comparison is that the URM/non-URM achievement gap in our data is 0.6 *SD*.

Table 2 provides our findings on the estimated variation in teacher effects on eighth-grade math test scores and in particular the distributions for teachers in different grades, which is how we assess the persistence/fade-out of teacher effects from grade to grade. In Panel A, we give descriptive statistics of our estimated teacher-effect coefficients, reminding the reader that these unadjusted estimates are the sum of the true effect and the estimation error. Note that the Jarque-Bera statistic rejects normality with a *p* value equaling 0 to all reported digits.

Panel B of Table 2 provides descriptive statistics of teacher effects corrected for measurement error. Two findings are immediately apparent. First, we ask about the effect of improving the entire distribution by increasing all value-added scores by 1 *SD* as in Chetty et al. (2014). We do this both for current teachers and for early-year teachers. Consistent with existing empirical evidence (e.g., Chetty et al., 2014; Rivkin et al., 2005), our estimates show that a 1 *SD* variation in teacher effectiveness is substantively large.[23] As suggested earlier, one metric is the difference in test scores between URM and non-URM students on eighth-grade math test scores, which is approximately 0.60 *SD* in our sample.[24] Our empirical Bayes estimate is that a 1 *SD* improvement in the effectiveness of a student's eighth-grade teacher would close half the URM/non-URM gap.

[23] Hanushek and Rivkin (2010) review a variety of studies and find that the average effect of a 1 *SD* change in teacher value-added is about .11 *SD* of student test score achievement, and Chetty et al. (2014) estimate an effect size of .14. Our estimates are larger, ranging from .11 to .34, depending on the grade level of the teacher, but are within the range of what has been found in the literature (e.g., Goldhaber et al., 2013; Nye et al., 2004).

[24] This is close to national estimates of this gap. For instance, in 2017, the Black-White gap in eighth-grade math test scores on the National Assessment of Educational Progress was about 0.83, and the Hispanic-White gap was 0.61 (Hansen et al., 2018).

In the Introduction, we referenced early work on the importance of teachers that indicated that having a good teacher for 5 years would make an enormous difference in student outcomes. The results in Table 2 support this early view. If we could arrange for students to have teachers with a 1 *SD* higher level of effectiveness in all five measured grades (adding the effects), we would see eighth-grade outcomes improved by 1 *SD*, which is almost twice the URM/non-URM gap. In traditional value-added estimates, questions arise as to whether the results are due in part to the sorting of students. This issue matters for some of our results (see discussion of robustness in Section 5). However, when we add across grades, the issue of sorting across the five grades is less critical. Effectively, we are dividing up eighth-grade test results among five teachers. If we assign part of the score to one teacher, we take those points away from another, leaving the five-grade total unaffected.

The second immediately apparent result is that while there is estimated fade-out, our empirical Bayes estimates show much more persistence than suggested by the prior literature on this issue. As we have noted previously, Jacob et al. (2010) find that only about a quarter of teacher contributions persist into the following year.[25] Similarly, Rothstein (2010) finds that a third of a third-grade teacher's effect persists for 2 years (Rothstein's Table VIII, first column). In contrast, comparing standard deviations, we find that the effects of seventh-grade teachers on eighth-grade test outcomes are about 60% as important as those of eighth-grade teachers. And even fourth-grade teachers are found to have an impact on eighth-grade tests that is about 32% as important as that of eighth-grade teachers. Note, however, that when we allow for a more flexible fitting of the teacher contribution (see discussion that follows), the estimate rises from 32% to 45%.

Given the evidence of non-normality, we also estimate the distribution of teacher effects corrected for measurement error using Delaigle's nonparametric method.[26] Delaigle-adjusted estimates are notably leptokurtic, more so in earlier grades. This plays a role in results reported in the remainder of this section. All the reported standard deviations are modestly larger. The increase is particularly noticeable for fourth-grade teachers, raising the estimated ratio of fourth- to eighth-grade standard deviation to 45%.

Next, we ask what the effect would be on long-term outcomes of replacing teachers below the 5th percentile of value-added with teachers with median value-added (the thought experiment posed in Hanushek, 2009). Formally, this asks the value of $F^{-1}(0.5) - \mathrm{E}\left(\tau \middle| \tau < F^{-1}(0.05)\right)$, where $F(\cdot)$ is the cumulative distribution function of teacher effects. We extend the experiment to also consider the loss from teachers in the upper tail of the effectiveness

[25] We replicate the technique of Jacob et al. (2010) using our data in the Appendix.

[26] Delaigle's method gives density estimates p_i on a grid at points $x_i, i = 1, \ldots, n$. We estimate the first moment of the distribution as $\mu = \sum_{i=1}^{n} x_i \left[\frac{p_i}{\sum_{i=1}^{n} p_i}\right]$. We estimate the kth central moment as $\mu_k = \sum_{i=1}^{n} (x_i - \mu)^k \left[\frac{p_i}{\sum_{i=1}^{n} p_i}\right]$.

distribution leaving a school and being replaced by a median teacher. Results are provided in Table 3. Looking first at the empirical Bayes estimates, we find, in line with Hanushek's findings, that the effect of replacing a teacher in the tail with a median teacher is large.[27] The effect of changing an eighth-grade teacher is a little larger than the observed gap in test scores between URM and non-URM students. The effect of changing early-year teachers is also large—about a third the effect of eighth-grade teachers for changing a fourth-grade teacher and half the effect for a fifth-grade teacher.

Table 3 presents analogous results using the Delaigle estimates. Given the high estimated kurtosis, it is not surprising that the results of replacing a teacher in the tail is larger than it appears in the empirical Bayes estimates. Moreover, the nonparametric estimates offer greater flexibility in allowing upper versus lower tail behavior to differ. The estimates of replacing a weak teacher with an average teacher are notably larger than the empirical Bayes estimates. Interestingly, the Delaigle results are quite different in the upper and lower tails. These estimates suggest that in eighth grade, replacing a teacher below the 5th percentile is more important than retaining a teacher above the 95th percentile, whereas the fourth-grade results go in the opposite direction.

Our results show that although there is teacher fade-out, the extent of fade-out is much smaller than what has been described in the earlier literature. This is useful to know from a macro policy perspective, but at ground level, an administrator may be interested in whether a specific early-year teacher who is successful in helping students achieve good current-year results is also likely to have contributed to those students' longer run outcomes. To address this issue, we repeated our estimates using earlier grades as the outcome variable. In other words, we regressed seventh-grade tests on indicators for teachers in Grades 4–7, sixth-grade tests on indicators for teachers in Grades 4–6, and so on.

In Table 4, we provide estimates of the correlations—the latent value-added performance over different grade outcomes[28]—between fourth-grade teachers' effects estimated on fourth-grade tests and the same teachers' effects estimated on later year tests.[29]

As is true with standard deviations, correlation coefficients need empirical Bayes adjustments for measurement error.

[27] Estimates for replacements in the lower versus upper tail differ slightly because the mean and median of the empirical Bayes shrunken estimates are slightly different.

[28] The estimate of fourth-grade teachers' effects on multiple-grade tests is based on the same sample of teachers to ensure the consistent sampling variability.

[29] To avoid the possibility that the same students contribute to the estimated teacher effects across multiple grades (creating a mechanical correlation), we use out-of-sample cohorts to estimate fourth-grade teacher effects on fourth-grade student outcomes. Specifically, we use the fourth-grade cohorts of students to which teachers are assigned in the 2015–16 to 2018–19 school years to estimate the effects on fourth-grade student tests, since the main results are based on cohorts from 2006–07 to 2014–15.

Equation (8) shows how we adjust the correlation by removing the noise estimates of performance/measurement error in this case,

$$corr(\tau^0_{j^{44}}, \tau^0_{j^{4g}}) = \frac{cov(\hat{\tau}_{j^{44}}, \hat{\tau}_{j^{4g}})}{\sqrt{var(\hat{\tau}_{j^{44}}) - \frac{1}{n}\sum \hat{\sigma}_{j^{44}}{}^2}\sqrt{var(\hat{\tau}_{j^{4g}}) - \frac{1}{n}\sum \hat{\sigma}_{j^{4g}}{}^2}} \quad (8)$$

where $\hat{\tau}_{j^{44}}$ is the estimate of fourth-grade teachers' effects on the fourth-grade test, $\hat{\tau}_{j^{4g}}$ is the estimated effects on *g*th-grade test, and σ is the standard errors of estimated effects for each. Consistent with the aforementioned earlier research on teacher fade-out, we see significant diminishment of estimated teacher effects as students move from grade to grade; indeed, by sixth grade, the correlation between teachers' impact on fourth-grade tests and their impact on sixth-grade tests is very close to 0.

5. Robustness Checks

We do three different types of robustness checks. The first is designed to assess the degree to which our approach to estimating value-added yields different estimates of teacher effectiveness than more traditional ways of estimating a teacher's contribution to student tests. The second is to check how sensitive our estimates of the distribution of teacher effects (and the correction for sampling error) are to the empirical Bayes approach defined in Equation 7. Finally, we assess the extent to which the sorting of students into different teacher classrooms, based on both observables and unobserved ability, may influence our findings on teacher fade-out.

Most value-added estimates are made by regressing current-year test scores on lagged test scores, control variables, and teacher indicators. Our estimation method is different, although not without precedent of closely related methods (Rothstein, 2010). A natural question arises: How different are the value-added estimates derived from our stacking teacher indicators across grades (Equation (6)) from those derived based on traditional year-to-year estimates of teacher contributions to student tests (Equation (9))?

$$Y_{ijt} = Y_{ijt-1}\beta_1 + X_{ijt}\beta_2 + I_{ijt}\tau_j^{VA} + \varepsilon_{ijt} \quad (9)$$

We assess this difference by correlating the effects from the two models and adjusting for sampling error, as described in Equation 8.

The correlations between these two different measures of value-added are provided in Table 5. In the case of eighth-grade teachers, both value-added estimates are based on the same student test outcomes, so it is not surprising that the adjusted correlations are quite high (more than 0.90). Interestingly, the correlation between a teacher's initial effects and the effects on

eighth-grade tests, while lower, are between 0.47 and 0.78, depending on the grades of comparison with eighth.[30] This finding suggests, consistent with the discussion in Section 2, that while some of the contributions that teachers make toward student learning are picked up by the test at the end of the grade in which they teach, teachers also have broader impacts on learning that contribute to later-grade (Grade 8) achievement. These contributions may well be in the form of math concepts that are not covered by specific grade tests or broader contributions to student learning. Recent evidence (Gershenson, 2016; Jackson, 2018; Kraft, 2019; Liu & Loeb, 2019), for instance, finds that teachers do make important contributions to various non-test student behaviors (e.g., attendance) that are not highly correlated with their effects on own-grade student test scores but may be quite important for longer term academic achievement.

As reported in Table 2, the estimates of the sampling-adjusted distribution of the teacher effects yields similar results whether we use a standard empirical Bayes adjustment or the Delaigle approach. But there are several other means by which research has recovered the estimated teacher effect. In particular, the permanent component of teacher quality can be recovered from the correlations of value-added estimates across years. As Kane and Staiger (2008) point out, assuming that the permanent component of value-added is constant across years, it is possible to recover the variance of teacher effects using Equation 8. In particular, the correlation between adjacent years of value-added is the covariance across years over the adjusted (for sampling error) standard deviation of the variance.[31] We obtain the sum of residuals and Grade *g*'s teacher effects from the value-added model and then average the total of both within a school year and create a lagged variable of this result to calculate the covariance between Year *t* and Year *t–1* for a teacher. In our estimation, for example, fourth grade, the residuals, and the fourth-grade teacher effect are calculated from Equation 6 by subtracting all the predicted values except the fourth-grade teacher effect from eighth-grade student scores. Thus, we obtained a variable of the sum of fourth-grade teacher effects and residuals. We take the square root of this value as the standard deviation of teacher effects in a particular grade.

In addition, we follow the Koedel et al. (2015) method to measure the overall R-squared (unadjusted) increase in the model with and without the inclusion of teacher indicators for each grade. This measurement reflects the incremental proportion of the total variance explained by the addition of teacher indicators.[32] We first calculate the R-squared of the full model (i.e., Equation 6 with all teacher-grade indicators), and then we omit the indicators for the teachers in

[30] These correlations are somewhat higher than those reported by Rothstein (2010), who presents similar evidence over a shorter grade span.

[31] Specifically, to obtain the adjusted standard deviation of the value-added estimates, we calculate the covariance of *g*th-grade value-added estimates at Years *t* and *t–1* in a teacher *j*: $\sqrt{cov(\hat{\tau}_{VA,jtg}, \hat{\tau}_{VA,jt-1g})}$. The denominator is the adjusted standard deviation of our estimates: $\sqrt{var(\hat{\tau}_{jg}) - \frac{1}{n}\sum \widehat{\sigma_{jg}}^2}$.

[32] It is a lower bound estimate of the contribution of Grade *g* teachers to explain eighth-grade student test scores, since the assignment of students to Grade *g* teachers may be correlated with their assignment to teachers in other grades that are included in the model.

a particular grade.[33] The difference between the full-model R-squared value and the model omitting the teacher indicators in Grade *g* is the measure of the proportion of the total variance explained by the teachers in Grade *g*. We take the square root of this value as the standard deviation of teacher value-added in that grade.

In **Table 6**, we report the estimated effect size of a 1 *SD* change in teacher value-added on eighth-grade test scores for each grade level of teachers, using the different methodologies described earlier. The first row simply replicates the empirical Bayes estimates (for sampling error) from Panel B of Table 2. The second row is the Delaigle estimates. Row 3 provides the estimates using the Kane and Staiger (2008) method, and Row 4 the square root of the increase of R-squared values between the full model and each grade's model using Koedel et al. (2015). Note that the estimates from each of these methods provide similar effect sizes of a 1 *SD* change in teacher value-added in each grade on eighth-grade student test achievement.

Finally, all estimates of value-added are imperfect, if only because, as noted previously, test-based value-added estimates only capture a slice of what teachers contribute to student achievement. A general acknowledgement of imperfection notwithstanding, there is one critique we think particularly relevant. Rothstein (2010) argues that value-added estimates may be biased due to the sorting of students into teacher classrooms based on unobserved attributes correlated with student achievement. We conducted a Monte Carlo experiment (described further on) to see the extent to which the sorting of students into teacher classrooms might bias our findings. Indeed, we do find evidence that the magnitude of the effect of early-grade teachers on later grade test achievement is influenced by the nature of student-teacher sorting, but we argue that the likely direction of the bias is to make our estimates of the importance of early-grade teachers conservative. The direction of bias hinges on the following question: Are students who are high performing based on unobserved factors more likely to be assigned to high-value-added teachers, or do low-performing students get the better teachers? While the latter might be preferable from an equity point of view, we strongly suspect that, due to the influence of parents and other social factors, the former case is more likely. Indeed, investigation of the distribution of observable teacher attributes across students tends to find that traditionally disadvantaged students (e.g., high-poverty students and URMs) are more likely to be assigned to less credentialed and experienced teachers, as well as teachers with lower value-added (Goldhaber et al., 2018; Isenberg et al., 2016; Mansfield, 2015; Sass et al., 2010).

Our Monte Carlo is simple but also designed to roughly follow the data structure in the state of Washington. We assume that students are only assigned to teachers in two grades ("early grade" and "late grade"), that teachers are observed in multiple school years, and that the number of years in which teachers are observed can vary (each teacher, on average, is observed two to

[33] For example, for a fourth-grade teacher, we calculate the R-squared of the model without the fourth-grade teacher indicator in Equation 6 and then take the difference between the full-model R-squared and this R-squared as the proportion of variance explained by fourth-grade teachers.

three times). Further, we assume that 40 students are assigned to each teacher-year track; that both early, Grade *t*, and late, Grade *t+1*, teachers have the same number of tracks; and that there are a total of nine student cohorts and 200 teacher-year tracks in each grade in the simulation. This results in a total of 8,000 student observations.[34]

The baseline test achievement for each student *i*, y_{i0}, is drawn from a standardized normal distribution, and we also assume that teachers' value-added, $\sigma_{tv,g}$, in early and late grades is normally distributed with mean 0 and a standard deviation of 0.15. We then assign students to early-grade teachers based on a sorting index S_{it}, which is equal to the sum of the baseline test score y_{i0} and an idiosyncratic shock (to represent the fact that students may be grouped by test performance but are not assigned to teachers based on test scores alone), ω_S.[35] Specifically, students are assigned to T_t groups in each grade, where T_t is the total number of teachers. There are three ways that students are assigned to teachers: (1) randomly, where the *T* groups of students are randomly distributed among the *T* teachers in each grade; (2) strong positive sorting, where the group of students with the highest average sorting index is assigned to the teacher with the highest value-added and so on; and (3) strong negative sorting, where the group of students with the lowest average sorting index is assigned to the teacher with the highest value-added and so on.

Early-grade student scores are generated after we assign early-grade teachers. This process contains three components: baseline test score with fade-out; teacher value-added of early grade; and a shock, as shown in Equation (10),

$$y_{it} = \gamma y_{i0} + \tau_{jt} + \omega_t \tag{10}$$

where γ is the parameter of persistence/fade-out (assumed to be 0.8), y_{it} is the scores of the early grade, τ_{jt} is the true teacher effectiveness in the early grade, and ω represents the student shock. This shock has normal distribution with mean 0 and a standard deviation of $1 - \gamma^2 - {\sigma_{tv,t}}^2$ for the sake of scaling to a unit variance. Late-grade teachers are assigned using the same mechanism, with S_{it+1} and early-grade student scores y_{it}. Then, we generate late-grade scores y_{it+1} based on the same logic.

We obtained the simulated student-teacher samples with the three different sorting mechanisms. Early- and late-grade teacher effectiveness, $\hat{\tau}_{j^g,g}$, where *g* is either *t* or *t+1*, are estimated by regressing late-grade student outcomes y_{it+1} on both early- and late-grade teacher

[34] The sample size in the Monte Carlo is much smaller than in the real data in order to limit computation time.
[35] The idiosyncratic shock ω_S is generated from a normal distribution with mean 0 and standardized deviation 5 to generate enough randomness.

indicators I_{jg} and baseline test scores y_{0i}. We obtained the teacher estimates and standard errors based on Equation (11) (which parallels Equation 6):

$$y_{it+1} = \sum_{j^{t+1} \in J^{t+1}} \hat{\tau}_{j^{t+1},t+1} I_{j^{t+1}} + \sum_{j^{t} \in J^{t}} \hat{\tau}_{j^{t},t} I_{j^{t}} + y_{0i} \quad (11)$$

Finally, we correct the standard deviation of estimated teacher effectiveness in Grade *g* by removing the sampling error.[36]

Table 7 reports the results of the Monte Carlo (recall that, by construction, the true distribution of teacher value-added in each grade is 0.15). Each panel of the table shows the estimated distribution (1 *SD*) of raw (i.e., not shrunken) teacher value-added (Column 1), the estimated distribution of shrunken estimates (Column 2), and the correlation between the shrunken estimates and the true teacher effects (Column 3). Panel A shows the findings for when students are randomly assigned to teachers, Panel B contains the findings with negative sorting, and Panel C displays those with positive sorting (again, this is the issue originally raised by Rothstein, 2010).

What is most relevant for interpreting our main results is what the simulation suggests about the ratio of early- to late-grade teacher standard deviations. Based on the true teacher value-added, this should be .8 (given the assumption of a .8 coefficient on the base-year test score in the data-generating process).

Panel A, in which student groupings are randomly assigned to teachers, shows the ratio of early- to late-grade teacher distributions of close to .8 for raw value-added and .645 for shrunken estimates of value-added. The shrunken estimates understate the true ratio due to overshrinkage associated with the teacher effects being imprecisely measured (Boyd et al., 2008).[37]

Panel B provides the estimates with negative sorting of students to teachers; again, we do not believe this is likely. In this case, both the raw and shrunken ratios of early- to late-grade teachers are overstated, by as much as 50% in the case of the shrunken values.

Panel C provides the findings for the concern about positive matching, raised by Rothstein (2010). Here, both the raw and shrunken ratios of early- to late-grade teachers are

[36] The adjusted standard deviation of our estimates is $\sqrt{var(\hat{\tau}_{jg}) - \frac{1}{n}\sum \widehat{\sigma_{jg}}^2}$, where $\widehat{\sigma_{jg}}$ is the estimated standard errors on teacher effectiveness.

[37] Increasing the number of students assigned to each classroom substantially improves the precision of the teacher value-added estimates, such that the shrunken parameters for early- and late-grade teachers converge to the true distributions of .12 and .15.

substantially understated; hence, we believe that our estimates of the true effects of early-grade teachers on eighth-grade student achievement (in Table 2) are conservative.

6. Conclusions

We partition eighth-grade test outcomes into components due to assignment to a particular eighth-grade teacher, to teachers in earlier years, to nonteacher factors, and to a random error. Our first conclusion is that whether a student is assigned to a particularly effective early teacher or not has an effect that is about half as large as the effect of being assigned to a particularly effective eighth-grade teacher. Thus, while there is some fade-out, a substantial portion of the impact of having a good early-year teacher continues much longer.

Our second conclusion is that the overall effect of having a good teacher is notable. Thinking of the Hanushek (2009) experiment, replacing the worst performing 5% of teachers with average teachers in Grades 4–8 would raise overall achievement by 2.5 *SD*. An important related finding is that the assumption of normality of the teacher value-added distribution masks the import of focusing on the tails of the distribution. This finding is particularly true for the most effective teachers, as the Delaigle estimates show that the effects of teachers at the upper end of the value-added distribution are even larger than those at the bottom. It implies that policymakers might wish to focus more attention on retaining the best teachers than dismissing the most ineffective.

At the same time, our results suggest a practical difficulty for administrators when thinking about retaining/replacing early-year teachers. We show that there is a large difference across early-year teachers in their contributions to later year outcomes. But administrators would not generally wish to wait until a current fourth-grade teacher's students took tests 4 years later to take a positive or negative action.

Perhaps a useful summary is that the findings of the earlier literature that teachers are incredibly important for student outcomes is sustained. There is, as an intervening literature suggests, some fade-out of teacher effects. But the fade-out is smaller than generally thought, and the effects of early-grade teachers are quite important.

References

Abowd, J. M., Creecy, R. H., & Kramarz, F. (2002). *Computing person and firm effects using linked longitudinal employer-employee data* (Technical Paper No. 2002-06). U.S. Census Bureau, Center for Economic Studies.

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

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*(4), 351–383.

Bartanen, B., & Husain, A. N. (2021). *Connected networks in principal value-added models* (EdWorkingPaper No. 21-397). https://doi.org/10.26300/5tjj-py73

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Measuring test measurement error: A general approach. *Journal of Educational and Behavioral Statistics, 38*(6), 629–663.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Overview of Measuring Effect Sizes: The Effect of Measurement Error. Brief 2*. National Center for Analysis of Longitudinal Data in Education Research.

Carrell, S. E., & West, J. E.. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409–432.

Cascio, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions* (NBER Working Paper No. 18038). National Bureau of Economic Research.

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences, 110*(43), 17176–17182.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633–2679.

Delaigle, A., Hall, P., & Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics, 36*(2), 665–685.

Delaigle, A., & Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli Journal, 14*(2), 562–579.

Gaure, S. (2013). OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis, 66*, 8–18.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*(2), 125-149.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109–127.

Gilraine, M., Gu, J., & McMillan, R. (2020). *A new method for estimating teacher value-added* (NBER Working Paper No. 27094). National Bureau of Economic Research.

Gilraine, M., Petronijevic, U., & Singleton, J. D. (2021). Horizontal differentiation and the policy effect of charter schools. *American Economic Journal: Economic Policy, 13*(3), 239-76.

Goldhaber, D., & Hannaway, J. (2009). Creating a New Teaching Profession. *Urban Institute Press.* 2100 M Street NW, Washington, DC 20037.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319), 589–612.

Goldhaber, D., Koedel, C., Özek, U., & Parsons, E., (2021). *Using longitudinal student mobility across schools and districts to identify at-risk students* (CALDER Policy Brief No. 25-0421). National Center for Analysis of Longitudinal Data in Education Research.

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has it always been this way? Tracing the evolution of teacher quality gaps in US public schools. *American Educational Research Journal, 55*(1), 171–201.

Goldhaber, D., & Startz, R. (2017). On the distribution of worker productivity: The case of teacher effectiveness and student achievement. *Statistics and Public Policy, 4*(1), 1–12.

Goldhaber, D., Theobald, R., & Fumia, D. (2018). *Teacher quality gaps and student outcomes: Assessing the association between teacher assignments and student math test scores and high school course taking* (Working Paper No. 185). National Center for Analysis of Longitudinal Data in Education Research (CALDER).

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.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher, 44*(5), 293-307.

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has it always been this way? Tracing the evolution of teacher quality gaps in US public schools. *American Educational Research Journal, 55*(1), 171-201.

Goldhader, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, 7(3), 199-208.

Hansen, M., Levesque, E., Valant, J., & Quintero, D. (2018). *The 2018 Brown Center report on American education: How well are American students learning?* The Brookings Institution.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). The Market for Teacher Quality. *NBER Working Paper,* (w11154).

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American economic review, 100*(2), 267-71.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3), 466–479.

Isenberg, E., Max, J., Gleason, P., & Deutsch, J. (2016). Do Low-Income Students Have Equal Access to Effective Teachers?. *Educational Evaluation and Policy Analysis*, 01623737211040511.

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics, 27*(2), 213–256.

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.

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

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of labor Economics*, *26*(1), 101-136.

Jochmans, K., & Weidner, M. (2019). Fixed-effect regressions on network data. *Econometrica, 87*(5), 1543–1560.

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

Kinsler, J. (2012). Beyond levels and growth estimating teacher value-added and its persistence. *Journal of Human Resources, 47*(3), 722–753.

Kinsler, J. (2016). Teacher complementarities in test score production: Evidence from primary school. *Journal of Labor Economics*, *34*(1), 29–61.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal, 48*(2), 361–386.

Koretz, Daniel M. (2009). *Measuring up: What educational testing really tells us.* Harvard University Press.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs?. *Education Finance and Policy*, 10(4), 508-534.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources, 54*(1), 1–36.

Leonhardt, D. (2010, July 28). The case for $320,000 kindergarten teachers. *The New York Times*. https://www.nytimes.com/2010/07/28/business/economy/28leonhardt.html

Linden, W. J., van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Springer Science & Business Media.

Liu, J., & Loeb, S. (2021). Engaging teachers measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources, 56*(2), 343-379.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67.

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics, 33*(3), 751-788.

Master, B., Loeb, S., & Wyckoff, J. (2017). More than content: The persistent cross-subject effects of English language arts teachers' instruction. *Educational Evaluation and Policy Analysis, 39*(3), 429–447.

Mihaly, K., McCaffrey, D., Sass, T. R., & Lockwood, J. R. (2013). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy, 8*(4), 459–493.

Nye, B., S. Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

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

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

Rose, J. (2021, January 26). *The grade-level expectations trap: How lockstep math lessons leave students behind*. *Education Next, 20*(2). https://www.educationnext.org/grade-level-expectations-trap-how-lockstep-math-lessons-leave-students-behind/

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

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2010). Value Added of Teachers in High-Poverty Schools and Lower-Poverty Schools. Working Paper 52. *National Center for Analysis of Longitudinal Data in Education Research*.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33.

Weeks, D. L., & Williams, D. R. (1964). A note on the determination of connectedness in an N-way cross classification. *Technometrics, 6*(3), 319–324.

Tables and Figures

Figure 1. Nonparametric Density of Teacher Effectiveness as Measured on Eighth-Grade Tests

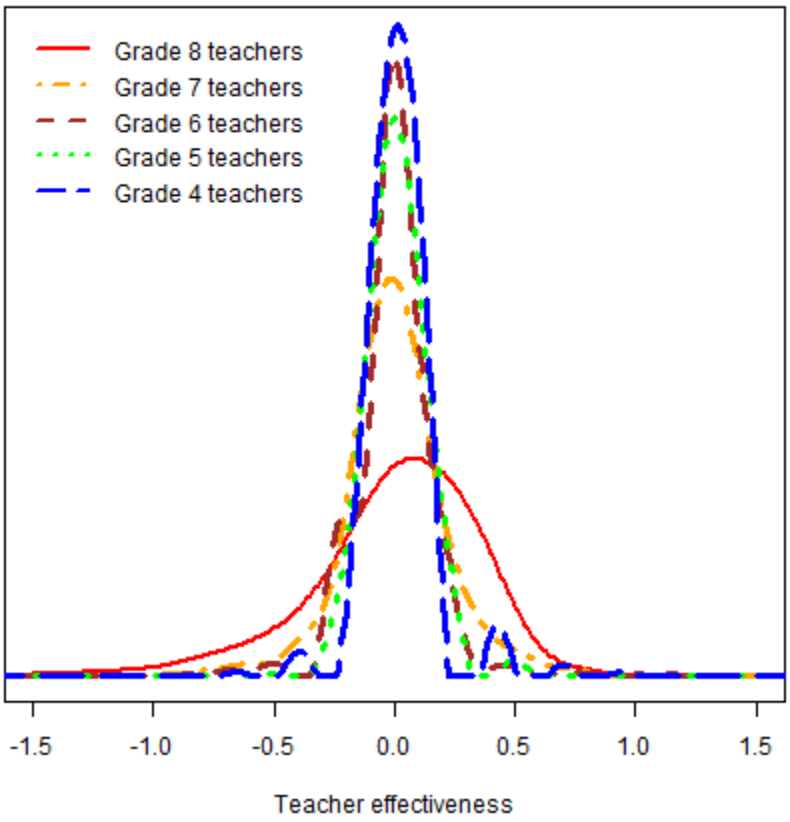

Figure 2. The Iceberg Problem and Assessing Teacher Contributions to Knowledge

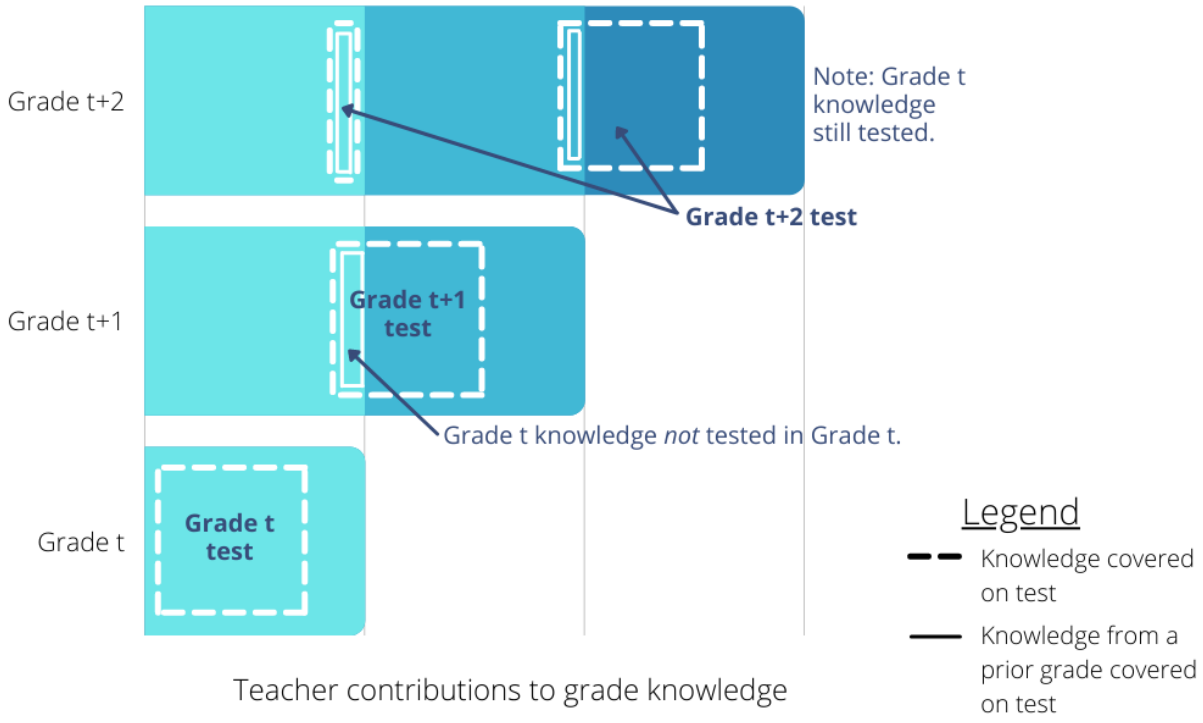

Figure 3. Example of Connected Teachers in Grades 4–8

Student ID	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Connected Group
1	A	H	N	R	W	
2	A	H	N	R	W	
3	A	I	N	R	W	
4	B	H	N	R	W	
5	B	I	N	S	W	
6	B	I	N	S	W	1
7	C	H	N	S	W	
8	C	I	N	S	W	
9	C	I	N	S	W	
10	D	J	N	S	W	
11	D	J	O	T	W	
12	D	J	O	T	W	
13	E	K	P	U	X	
14	E	K	P	U	X	
15	E	L	P	U	X	
16	F	L	P	U	X	2
17	F	L	P	U	X	
18	F	L	P	U	X	
19	G	M	Q	V	Y	3
20	G	M	Q	V	Z	

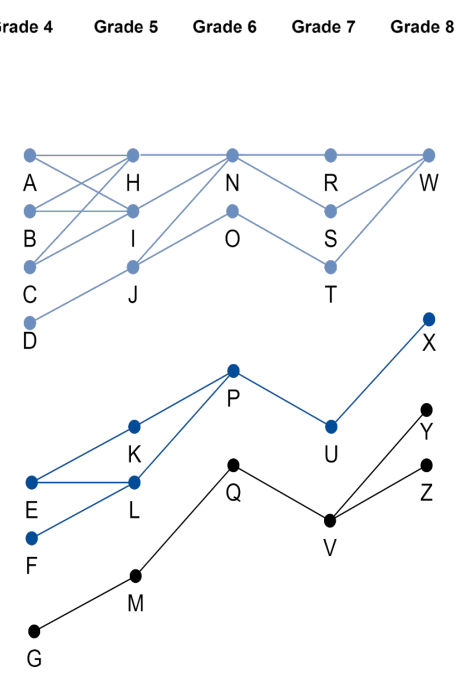

Table 1. Descriptive Statistics

	(1) Largest Connected Group	(2) Other Connected Groups	(3) Difference
URM student	0.251	0.315	-0.063***
Female	0.497	0.494	0.003
Asian/Pacific Islander	0.109	0.132	–0.023***
Third-grade FRL status	0.395	0.494	–0.099***
Third-grade special education	0.098	0.124	–0.026***
Third-grade English learner	0.090	0.118	–0.028***
Third-grade gifted	0.039	0.033	0.005***
Third-grade math test	0.065	–0.101	0.167***
	(0.978)	(1.025)	
Third-grade ELA test	0.068	–0.106	0.174***
	(0.976)	(1.028)	
Eighth-grade math test	0.236	0.043	0.193***
	(0.931)	(0.977)	
Number of unique students	173,858	112,116	

Note. Column 1 is calculated based on the largest connected group that is obtained from partitioning the dataset using Weeks & Williams (1964) algorithm. Column 2 is the summary statistics on the data excluding the largest connected group. In Panel A, we summarized student time-invariant variables, including those of underrepresented minority (URM) status, gender, and Asian/Pacific Islander indicator. Also, we have baseline controls for students including third-grade math and English language arts (ELA) scores (standardized to mean 0 and unit standard deviation prior to partitioning data into connected components), third-grade special education status, third-grade English learner status, and third-grade gifted indicator. In Panel B, we reported eighth-grade student math outcomes for both the largest connected group and other groups. The numbers of unique students and teachers in Column 1 are counted within the largest connected group. The number of unique students and teachers in Column 2 are those students and teachers who are not included in the largest connected group. Column 3 shows the differences between the largest connected group and others. Values are obtained from *t* test; all except gender are significantly different. Missing teachers are identified as missing in each grade and are not included/summarized in this table. FRL = free or reduced-priced lunch.

* $p < 0.05$
** $p < .01$
*** $p < 0.001$

Table 2. Estimates of Teacher Effects on Eighth-Grade Test Scores

	Grade				
Panel A: Raw estimates	**4**	**5**	**6**	**7**	**8**
SD	0.319	0.346	0.354	0.349	0.447
Skewness	0.043	–0.327	0.178	0.004	–0.206
Kurtosis	6.144	6.670	4.331	2.862	1.964
Jarque-Bera *p* value	0	0	0	0	0
Number of teachers	6,629	5,340	3,427	2,898	3,060
Average number of students per teacher in sample	18.3	18.3	28.0	45.5	56.6
Panel B: Adjusted estimates					
Empirical Bayes *SD*	0.111	0.176	0.188	0.209	0.337
Empirical Bayes shrunken kurtosis	0.585	0.573	1.454	1.081	0.177
Delaigle-adjusted *SD*	0.162	0.190	0.192	0.225	0.360
Delaigle-adjusted Kurtosis	37.1	52.5	21.8	6.2	5.6

Note. The table includes descriptive statistics for coefficient indicator variables for teachers.

Table 3. Estimates of Replacing Teachers in the Tail of the Effectiveness Distribution on Eighth-Grade Test Scores

	Grade				
	4	**5**	**6**	**7**	**8**
Replace below 5th percentile with median teacher, empirical Bayes estimates	0.229	0.362	0.386	0.421	0.685
Replace above 95th percentile with median teacher, empirical Bayes estimates	0.229	0.364	0.389	0.440	0.707
Replace below 5th percentile with median teacher, Delaigle estimates	0.294	0.395	0.403	0.526	0.970
Replace above 95th percentile with median teacher, Delaigle estimates	0.406	0.342	0.420	0.523	0.566

Note. See article text for calculation method.

Table 4. Adjusted Correlations of Fourth-Grade Teacher Effects with Different Outcome Grades

Outcome	Correlation
Fifth-grade test	0.22
Sixth-grade test	0.07
Seventh-grade test	0.01
Eighth-grade test	–0.10

Note. The outcome test is estimated by Equation 6, with the number of teacher indicators varying according to the outcome grade in question (e.g., the sixth-grade test has indicators for fourth-, fifth-, and sixth-grade teachers). Column 2 shows the correlation between fourth-grade teachers' effect on fourth-grade test scores and fourth-grade teachers' effect on each outcome grade test score. Fourth-grade teachers' effect on fourth-grade test scores are estimated using out-of-sample cohorts from 2016 to 2019. This correlation has been adjusted by using the empirical Bayes adjustment as described in Equation 8.

Table 5. Correlations of Our Value-Added Estimates with Traditional Estimates

Grade	Correlation
4	0.49
5	0.47
6	0.52
7	0.78
8	0.94

Note. Value-added estimates are based on the largest connected group sample that includes student cohorts from 2007 to 2019. Column (1) of the table shows the grade in which traditional value-added models (defined by Equation 9) has been estimated and correlated with the teacher-grade indicators models described in Equation 8. Column 2 shows the adjusted correlation that removed sample errors in the estimated standard deviation that is the square root of covariances in mean residuals across a classroom for each teacher. Each classroom is paired with a randomly chosen classroom of the same teacher to estimate the covariances.

Table 6. Different Methods of Deriving Teacher Effect Size Estimates

	1 *SD* change in teacher value-added on student tests			
	Grade			
	4	**5**	**6**	**7**
Empirical Bayes	0.11	0.18	0.19	0.21
Delaigle	0.16	0.19	0.19	0.23
Kane & Staiger	0.14	0.15	0.17	0.21
Koedel et al.	0.13	0.12	0.11	0.14

Note. Column 1 shows the adjusted empirical Bayes standard deviations. Column 2 has the adjusted empirical Bayes standard deviations with clustered standard errors on eighth-grade classroom level. Column 3 shows the square roots of the covariance of the permanent component of teacher quality in Years *t* and *t–1* using Kane and Staiger (2008). Column 4 displays square roots of the increase of the R-squared difference with and without the inclusion of a particular-grade teacher indicator using Koedel et al. (2015).

Table 7. Monte Carlo Results on Teacher Value-Added Distribution

Panel A: No Sorting of Students to Teachers			
	1 *SD* Raw Value-added	1 *SD* Shrunken Value-added	Correlation of True Value-added and Shrunken
Early-grade teachers	.138	.091	.852
Late-grade teachers	.174	.141	.849
Ratio of early to late *SD*	.793	.645	
Panel B: Negative Sorting of Students to Teachers			
Early-grade teachers	.135	.085	.841
Late-grade teachers	.116	.056	.774
Ratio of early to late *SD*	1.164	1.518	
Panel C: Positive Sorting of Students to Teachers			
Early-grade teachers	.135	.085	.843
Late-grade teachers	.212	.184	.938
Ratio of early to late *SD*	.637	.462	

Appendix A

Replication of Jacob et al. (2010) 2SLS Persistence Model

Here we describe our replication of the Jacob et al. (2010) measure of teacher persistence, using the largest sample of connected teachers. Specifically, we first obtain the ordinary least squares estimate of teacher persistence from a regression model that includes student demographics (X_i) and school year variables:

$$Y_{ijg} = \alpha_i + \beta_{ols} Y_{ij\,g-1} + \beta_2 X_i + \epsilon_{ijg}$$

in which students' achievement Y_{ijg} and $Y_{ij\,g-1}$ represent the test scores of students *i* matched with Teacher *j* at Grade *g* and Grade *g–1*, respectively. We construct similar estimate for β_{VA} and β_{LR} from these two stage least squares regressions:

$$Y_{ijg} = \alpha_i + \beta_{LR} Y_{ij\,g-1} + \beta_2 X_i + \epsilon_{ijg}$$
$$Y_{ijg-1} = \alpha_i + \beta_z Y_{ij\,g-2} + \epsilon_{ijg-1}$$

where $Y_{ij\,g-2}$ is the twice-lagged student achievement and β_{LR} is the persistence of long-run knowledge. And we use teacher j's effectiveness $\tau_{ij\,g-1}$, i.e. value-added estimates, from Grade *g*–1 as an instrumental variable for the achievement of Grade *g*–1, we then acquire estimate of β_{VA} from a regression model:

$$Y_{ijg} = \alpha_i + \beta_{VA} Y_{ij\,g-1} + \beta_2 X_i + \epsilon_{ijg}$$
$$Y_{ijg-1} = \alpha_i + \beta_v \tau_{ij\,g-1} + \epsilon_{ijg-1}$$

We also extend 1-year persistence model to 2-year persistence model according to Jacob et al. (2010) by taking one more lag on student achievements and using teacher value-added as instrumental variable from one more year earlier. These three measurements of persistence are interpreted, respectively, as the fraction of variance in total knowledge attributable to long-run knowledge, β_{LR}, and the fraction of long-run knowledge related to teacher effects, β_{VA}.

In Table A1, we report the 1- and 2-year persistence estimates on math achievement, i.e., β_{VA} using the procedure from Jacob et al. (2010) model on Grades 4–6 and 4–8.[38] The Jacob et al. estimates of 1- and 2-year persistence are presented in Panel A.[39]

Our lagged score and value-added instrumental variable approaches for Grades 4–6 is reported in Panel B. These estimates of long-run persistence with lagged score instrument for both 1- and 2-year persistence are very similar to Jacob et al. (2010) estimates; the reported Jacob et al. 1-year value-added persistence of knowledge is 0.27 in North Carolina in Grades 4–6 (p. 929). Our 1-year estimate with value-added instrument on the largest connected sample of

[38] We report the results separately for these grades to be consistent with the Jacob et al. (2010) sample (Grades 4–6) and to show that the results do not vary significantly for the later grades in our full sample.

[39] We also reproduce findings using the largest connected sample and a different approach advanced by Kinsler (2012). Our estimates of persistence using Kinsler's approach is 0.51, which is somewhat higher than that reported in Kinsler (0.375).

teachers in Grades 4–6 is slightly smaller, 0.18, and about 0.07 in 2-year value-added persistence estimates. In Panel C, we extend the persistence estimates to Grades 4–8 using the same specification. We have slightly higher long-run persistence estimates using lagged score IV approaches. The estimated 1-year value-added persistence of teachers in Grades 4–8 is similar, around 0.26. The 2-year value-added persistence of teacher is smaller, 0.08.

Table A1. Estimates of the Persistence of Math Achievement

	One Year		Two Year	
	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$	$\hat{\beta}_{LR}$	$\hat{\beta}_{VA}$
Panel A: North Carolina Grades 4–6 (Jacob et al., 2010)				
Prior-year achievement coefficient	0.95***	0.27***	0.87***	0.16***
	(0.001)	(0.01)	(0.001)	(0.008)
Panel B: Washington State Grades 4–6				
Prior-year achievement coefficient	0.96***	0.18***	0.89***	0.07***
	(0.004)	(0.021)	(0.007)	(0.027)
Panel C: Washington State Grades 4–8				
Prior-year achievement coefficient	0.97***	0.26***	0.91***	0.08***
	(0.002)	(0.015)	(0.003)	(0.016)

* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$

Note. Results shown here are based on the largest connected sample with value-added estimates. For teacher value-added persistence, we estimate the value-added of a student's teacher from Jacob et al. (2010) and then estimate the persistence of teacher value-added using equations in Appendix A. Also, the long-run persistence is estimated similarly using lagged test scores as instrumental variables. We applied the Jacob et al. method on North Carolina from 2007 to 2017 and obtained similar results comparing to their estimates. End-of-grade student achievement is standardized within grade and school year. We only keep those students who can be matched with a single teacher in a single grade year using the preferred matching protocol. Standard errors are obtained from 2SLS ordinary least squares.