Classroom Assignment Policies and Implications for Teacher Value-Added Estimation

Hedvig Horváth*

January 12, 2015

Please click here for the most recent version.

Abstract

Value-added measures of teacher quality may be useful tools to help manage the teacher workforce and improve the efficiency of schools. But this requires that they be reliable estimates of teacher effectiveness. Because value-added scores are based on observational data, they may be biased by systematic patterns in the assignments of students to teachers. Whether assignment processes permit unbiased estimation of teacher value-added is a matter of great dispute. In this paper, I loosen the assumption from past research that assignment processes are identical at all schools. I develop a method for measuring the heterogeneity in school-level assignment practices, and I leverage the variation in practices to learn about the magnitude of biases in teacher value-added estimates. I show that about 60% of elementary schools in North Carolina systematically sort students with higher and lower scores on previous year's tests to different classes - a pattern of student "tracking". About half of these schools also allocate the classes of high and low achievers to the same teachers year after year – a pattern of "matching" teachers to certain students. Biases in value-added estimates are most likely in these matching schools, and least likely in schools that are neither tracking nor matching. Using data on teachers who move between the two types of schools, I document strong evidence of systematic sorting and substantial biases in value-added measures. Importantly, these biases are negatively correlated with teachers' true effects, so would not be detected by prior estimates of value-added. Overall, I conclude that the quality of value-added assessments is likely to depend on the nature of the student-teacher allocation process used at specific schools or school systems.

*Ph.D. Candidate, Department of Economics, UC Berkeley, hedvigh@econ.berkeley.edu. Words can barely express how grateful I am to David Card and Jesse Rothstein for being so generous with their guidance, time and encouragement through this project and throughout my Berkeley years. I am also indebted to Patrick Kline and Christopher Walters for their insights and suggestions, to Attila Lindner, John Mondragon, Carl Nadler, Eva Cheng, Tadeja Gračner, Jamie McCasland, Ana Rocca, David Silver and Moises Yi for helpful discussions and feedback. Seminar participants at UC Berkeley provided valuable comments. All remaining errors are solely mine.

1 Introduction

After many years of unsuccessful search for observable characteristics that can identify more or less effective teachers, a new body of research suggests that relatively simple measures of teacher effectiveness – constructed from student test scores that are routinely administered in virtually all public schools – can be used to classify teachers and help manage the teacher workforce.[1] Indeed, a number of recent papers have argued that value-added measures are reliable enough to serve as a basis for firing and promotion decisions, and for allocating financial rewards.[2]

Despite the enthusiasm of many researchers, a number of criticisms have been raised of value-added assessment procedures. One concern is that value-added measurements may create incentives for teachers to "teach to the test". Answering this concern, Chetty et al. [2014b] show that high value-added teachers also have long-run impacts on non-test related outcomes such as college enrollment and earnings.[3] A second concern is that value-added scores may not be fair, unbiased estimates of teachers' effectiveness. Value-added estimates are coefficients from simple OLS regression models on observational data, and may be biased if students are not randomly assigned to classrooms, and if the model controls – most importantly, the student's prior year scores – are not sufficient to absorb this non-randomness. Economists are generally skeptical of "causal" interpretations of such purely observational estimates when there is no *a priori* reason to expect random assignment.

In fact, there is considerable anecdotal evidence that the assignment process of students to teachers in many schools is far from "random". Many parents have had the experience of requesting a preferred teacher for their child. In many schools, teachers confer to decide

[1] Earliest examples include Rockoff [2004]; Nye et al. [2004]; Rivkin et al. [2005].

[2] See e.g. Aaronson et al. [2007] or "Big Study Links Good Teachers to Lasting Gain" in New York Times, January 6, 2012 (source: http://www.nytimes.com/2012/01/06/education/big-study-links-good-teachers-to-lasting-gain.html?pagewanted=all&_r=0) and its continuation at "Can a Few Years' Data Reveal Bad Teachers?" (New York Times, Room for Debate, January 16, 2012, source: http://www.nytimes.com/roomfordebate/2012/01/16/can-a-few-years-data-reveal-bad-teachers), and elsewhere in the blogosphere.

[3] Rothstein [2014] argues that the magnitude and significance of these results are largely sensitive to the inclusion of control variables. On the same point, teaching to the test may be a reason for the moderately strong correlation between different test score based value-added measures and teacher evaluations from principal/student/other observer ratings. See more in Harris and Sass [2007]; Jacob and Lefgren [2008]; Kane et al. [2013] and Rothstein and Mathis [2013].

on classroom assignments for the incoming students; in others, the decision is made by the principal (in both cases with input from parents). The process rarely appears on its face to resemble simple random assignment.

I show below that even in elementary grades, many schools systematically sort higher and lower achieving students into similar classes, and systematically assign classes of higher or lower achieving students to the same teachers year after year. Systematic sorting of weaker and stronger students to different classes may be helpful in allowing teachers to target their lessons (Duflo et al. [2011]). Allocating certain teachers to the lower-level classes, and others to the higher-level classes may allow teachers to specialize, or exploit their own comparative advantage. It may also serve as a mechanism that allows principals to reward certain teachers, or punish others.

A natural question is whether the systematic patterns of classroom assignments create bias in value-added scores. Research on this question is inconclusive. Recent papers such as Chetty et al. [2014a] or Bacher-Hicks et al. [2014] argue, based on quasi-experimental analyses of teacher switches, that sorting to classrooms based on unobservable characteristics is so small that the bias it introduces to value-added estimates is negligible. On the other hand, Rothstein [2014] questions the validity of the teacher-switching quasi-experiment, and in line with earlier works (Rothstein [2009], Rothstein [2010]), he maintains that the estimates may carry a nontrivial teacher-level bias.

Existing estimates of bias in value-added scores have two important features in common. First, all are predicated on the implicit assumption that any bias in a teacher's value-added score is uncorrelated with her true effectiveness. This assumption is unfounded, and may not hold in practice. Second, they treat the presence or absence of bias and its magnitude, as constant across a whole school system, and have nothing to say about any variation across schools with potentially quite heterogeneous assignment practices.

This paper contributes to the literature by developing a new framework for assessing the importance of nonrandom classroom assignment at the school level, in North Carolina.[4] I

[4]I focus on assignment to classrooms within schools. Assignment of students to schools is clearly not random,

define two kinds of nonrandom classroom assignment practices, tracking and matching, that have often been lumped together in past work but that have different implications for value-added analysis. A school tracks if it assigns students to classrooms on the basis of prior test scores (or other characteristics correlated with them). A school matches if the same teachers tend to get the high-prior-score classrooms year after year.

Building on a measure introduced by Clotfelter et al. [2006], I construct measures of tracking and matching based on variance decompositions of prior test scores and other predetermined student characteristics into components that vary across teachers, across classrooms (years) within teachers, and within classrooms. I explicitly consider the possibility that school-level measures, which can have low power, may misclassify schools. The test for tracking has great precision, though the power of the test to detect matching is lower.

I estimate that 60% of North Carolina elementary schools engage in tracking, and about half of these schools also engage in matching. At another 30%, classroom assignments appear to be purely random.

My measures of nonrandom classroom assignment is purely based on observable characteristics of students (e.g. lagged test score) that can be controlled for in value-added models. However, similarly to Rothstein [2009] or Altonji et al. [2005] and Oster [2014] in other contexts, I argue that under most reasonable scenarios, sorting based on observables signals that sorting on unobservables is likely as well, and thus at least suggests the potential for bias in value-added scores. In support of this, I find that at many tracking & matching schools parental education – generally not controlled for in value-added models – predicts teacher assignments conditional on the lagged test score. This points to the possibility that teacher value-added scores may be biased for teachers at tracking & matching schools, where there is less reason to expect bias at schools that neither track nor match.

I next turn to quantifying the bias at the tracking & matching schools. My strategy for this exploits teacher movements between schools, either within type (matching or not) or between.

and represents another source of potential bias in value-added scores, but is not considered here. (Clotfelter et al. [2006])

I show that the covariance of a teacher's value-added score from one year to the next varies between stayers and different types of movers, and that these covariances, along with the across-teacher variance of value-added estimates at each type of school, identify the magnitude of any biases and their correlation with teachers' true effects.

Method of moments estimates indicate that the variance of biases at matching schools is substantial, but that these biases are strongly negatively correlated with teachers' true effects. Immediate implications are that how well teacher value-added scores predict student achievement on average may be context-dependent: they may do quite a good job in randomly assigning schools but less so in matching ones.

The correlation between the bias and the true teacher effects has not previously been estimated, but has important implications. In particular, it can help to reconcile my results with those of other studies that find little or no evidence of bias. As discussed by Chetty et al. [2014a], existing strategies for measuring bias (the teacher switching quasi-experiment used by Chetty et al. [2014a] and Bacher-Hicks et al. [2014], as well as experimental methods used by Kane and Staiger [2008] and Kane et al. [2013]) have zero power in the "knife-edge" case where the covariance between bias and teachers' true effects equals minus the variance of the bias. My point estimates are quite close to this value, which always lies within the confidence interval, implying that these past estimates are likely to have substantially understated the true magnitude of biases. My results are thus highly relevant for policy: While Chetty et al. [2014a] and others' estimates of "forecast bias" tell us that value-added models commonly used yield approximately unbiased predictions of the change in average student achievement following a change in the teaching pool at a school (subject to Rothstein [2014]'s critique), they are not informative about the use of these models to evaluate individual teachers. This use appears much more tenuous and subject to quantitatively important biases.

This paper closely relates to two recent papers. Clotfelter et al. [2006] examine heterogeneity in school matching practices as they relate to teachers' observable characteristics. Kalogrides et al. [2013] estimate correlations between teacher observables and students' lagged scores, within school-grade-year cells. Both papers find evidence for some sorts of matching, and both

highlight the potential implications on value-added estimation.

Two other papers pursue that issue. Alzen and Domingue [2013] classify schools as tracking and non-tracking, and investigate differences in teacher value-added between the two types. They do not find evidence of any differences, though their sample is quite small. Dieterle et al. [2013] distinguish between tracking and matching, as I do, and investigate the sensitivity of different value-added model specifications to these types of nonrandom classroom assignments.

My analysis builds on this work in a number of ways. First, I use a simpler, more flexible measure of tracking and matching than have past authors. Where Dieterle et al. [2013], like Kalogrides et al. [2013] and Clotfelter et al. [2006], can detect only matching based on teacher observables, my measure detects other varieties of matching (such as that based on a teacher's true effectiveness). Second, I explicitly account for the potentially low power of the tests, which turns out to be important for the analysis of matching – many schools that in fact match will not be detected as such in the available data. Third, and most important, I build on the matching analysis to identify key moments of the distribution of bias in value-added scores, rather than simply pointing to the possibility for such bias or, as in Dieterle et al. [2013], documenting differences in scores obtained from different specifications of the value-added model.

The paper is organized as follows. Section 2 lays out why nonrandom classroom assignment is problematic in teacher value-added estimation, while in Section 3 I describe the dataset. Section 4 outlines the strategy and results of identifying schools that engage in some nonrandom classroom assignment practices, either tracking or matching. In Section 5, I briefly describe the method I use to estimate teacher value-added. Furthermore, I set up the structure of teacher value-added that school-level classroom assignment practices imply and show how that structure can be used to pin down the magnitude of the bias in value-added estimates. The last subsection of Section 5 contains the main results and a discussion, while Section 6 concludes.

2 Framework

I adopt the framework of value-added models (VAMs) to develop a new strategy to assess whether they yield unbiased estimates of teacher effectiveness.

Value-added models are versions of education production functions (Hanushek [1971]), which focus on isolating the effect of one specific input (mostly schools or teachers) on short-run testing outcomes, while assuming that the influence of other or earlier inputs are absorbed by prior achievement and other controls. I settle with the common specification of additively separable inputs, with the goal to estimate teacher effects. Assume that the true data generating process for the end-of-grade test score of student i in year t is

$$y_{it} = X_{it}\beta + Z_{it}\rho + \nu_{it}, \tag{1}$$

where

$$\nu_{it} = \mu_j + \theta_{jt} + \varepsilon_{it}, \tag{2}$$

as in Kane and Staiger [2008]. X_{it} are observable student characteristics, including prior achievement. They are also meant to capture school-grade-year level determinants of achievement, such as the principal or the facilities in school.[5] Z_{it} consists of student characteristics that are unobserved to the econometrician, but may be observable for the school principal, who can thus use both X_{it} and Z_{it} to assign students to classrooms. When thinking about X and Z, I will primarily have lagged scores (X) and parental involvement (Z) in mind. Therefore, and without loss of generality, I will assume that both factors in X and Z increase test scores, that is, $\beta > 0$ and $\rho > 0$.

Most importantly in the specification above, ν_{it} can be decomposed as (2), where $J(i,t) = j$ denotes the teacher assignment of student i in year t, *within the school*. μ_j's, the time-invariant teacher effects, are the parameters of interest. θ_{jt} captures classroom-level influences,

[5] In the analysis below, in some specifications, I either include school-grade-year fixed effects or school-grade-year peer characteristics to account for such factors.

such as year-to-year variation in within-classroom peer interactions. It can also reflect student assignments: If the group of students assigned to teacher j in year t is higher performing than would be expected based on the characteristics X_{it} and Z_{it}, this will manifest as a positive θ_{jt}. ε_{it} is student-level unobserved heterogeneity, net of any component that is common in the whole classroom and therefore picked up by θ_{jt}.

As Z_{it} is unobserved, it cannot be included in the estimated version of eq. (1), so the researcher will end up estimating the partially controlled equation,

$$y_{it} = X_{it}\tilde{\beta} + \tilde{\nu}_{it}, \tag{3}$$

where $\tilde{\beta} = \beta + \kappa\rho$, and κ is the projection coefficient of Z on X, which captures the correlation between Z and X. In my interpretation of X and Z, it is natural to assume $\kappa > 0$, and the residual, $\tilde{Z}_{it}$ can be thought of as the inherent assertiveness of parents, independently of the child's lagged scores. Then, since $\tilde{Z}_{it} = Z_{it} - \kappa X_{it}$,

$$\tilde{\nu}_{it} = \tilde{Z}_{it}\rho + \nu_{it} = \mu_j + \tilde{\theta}_{jt} + \tilde{\varepsilon}_{it},$$

where

$$\tilde{\theta}_{jt} = \overline{\tilde{Z}}_{jt}\rho + \theta_{jt} \tag{4}$$

and $\tilde{\varepsilon}_{it} = \left(\tilde{Z}_{it} - \overline{\tilde{Z}}_{jt}\right)\rho + \varepsilon_{it}$.[6]

The assumption needed to identify μ_j is that there is no systematic variation in $\tilde{\theta}_{jt}$ across teachers. As classrooms are nested within teachers, it is not possible to estimate teacher effects with fixed effects for classrooms. Accordingly, value-added specifications typically treat $\tilde{\theta}_{jt}$ as part of the error term. When this is done, μ_j is seen as the component of the classroom mean that is common across years for the same teacher, and $\tilde{\theta}_{jt}$ as the annual deviation from this. The implicit assumption is that the components of the classroom-level outcome other than the teacher's causal effect – including, in particular, the component attributable to unobserved

[6] Averaging of $\tilde{Z}$ takes place at the teacher-year level.

differences in students – are independent across classrooms within teachers.[7] A central question in value-added estimation is whether this is correct, or whether the resulting estimates of μ_j are contaminated by persistent components of student assignments that are not captured by the control variables X_{it}.

Schools may very well be heterogenous in their classroom assignment practices. To formalize this idea, I build on Rothstein [2009]'s framework. I suppress subscripts for easier readability. Without loss of generality, imagine there are two teachers at a school where the assignment of students to these teachers depends on an index composed of X and Z, and a purely random component η:

$$T = X\pi_1 + Z\pi_2 + \eta, \tag{5}$$

where $T \in \{0, 1\}$, as in a linear probability model. π_1 and π_2 capture the extent of sorting students to teachers based on observables and unobservables. Positive values of π_1 and π_2 can be interpreted as students with X and Z values that predict higher test achievement being more likely to get assigned to teacher 1. If teacher 1 is the better (higher μ) teacher, then $\pi_1 > 0$ means positive sorting on observables, while $\pi_2 > 0$ means positive sorting on unobservables.

Note that sorting parameters π_1 and π_2 play an important role in determining whether value-added scores are estimated with a bias or not. Some schools may randomly assign their students to teachers by setting $\pi_1 = \pi_2 = 0$, in which case value-added estimates obtained from specification (3) are unbiased.[8] However, in the more likely case when schools do sort and set $\pi_1 > 0$ and more importantly, $\pi_2 > 0$, value-added estimates will be biased.[9]

Below, I document that at some schools students' observables, X_{it}, are systematically sorted across classrooms, but that there is no sign of persistence in the type of classrooms that a particular teacher is assigned. This practice, which I call "tracking", appears consistent with

[7] One could, in addition, include classroom-level peer characteristics to make this assumption more credible. I do this in some specifications below.

[8] In fact, in the special case when not just $\pi_2 = 0$ but $\pi_1 = 0$ as well, value-added scores from the uncontrolled specification, $y_{it} = \mu_j + \mathring{\theta}_{jt} + \mathring{\varepsilon}_{it}$ are also unbiased. See the formal derivation for both specifications in the general case of J teachers in Appendix A.1.

[9] Theoretically, $\pi_1, \pi_2 \neq 0$ is the problem. I pick $\pi_1, \pi_2 > 0$ for illustrative purposes, because that case is more in line with the anecdotes one hears.

the above value-added specification. Insofar as the students are grouped into classrooms based on unobservables as well as observables, this will be reflected in the $\tilde{\theta}_{jt}$ error term. As long as there is no systematic variation in $\tilde{\theta}_{jt}$ *within teachers, across years*, this does not pose a problem in estimating μ_j in an unbiased way. At other schools, however, the same teachers are assigned students with high prior test scores year after year, while others are consistently assigned low prior score students. I refer to this practice as "matching". It is less clear whether value-added specifications can recover teachers' causal effects in these schools. If the observable controls in the model can fully absorb all of the across-teacher differences in the students assigned, the model will be successful. Otherwise, value-added estimates for individual teachers will be biased.[10]

In Section 4, I statistically distinguish between schools at which student observables, particularly lagged scores, are systematically sorted across classrooms and teachers and those where I am unable to reject the hypothesis that students are randomly assigned to classrooms, and therefore, also to teachers. Since Z is unobservable, rather than by eq. (5), the assignment process in practice can be characterized by

$$T = X\left(\pi_1 + \kappa\pi_2\right) + \tilde{\eta}, \tag{6}$$

where $\tilde{\eta} = \tilde{Z}\pi_2 + \eta$. Therefore, the distinction between random and systematically sorting schools based on X is formally a statement about $\pi_1 + \kappa\pi_2$ being 0 or not. The most natural interpretation of this is, however, to assume that the first group, the "tracking & matching" schools set $\pi_1 \neq 0$, while the second group, the random schools set both $\pi_1 = \pi_2 = 0$. Under this interpretation, there will be no bias in the value-added scores at the random schools, independent of what controls we use in the value-added specification. There is a knife-edge case of parameter values, $\pi_1 = -\kappa\pi_2 \neq 0$, when schools appear to be randomly assigning,

[10]Importantly, it is not necessary for *all* of the factors that influence teacher assignments – which will typically include hard-to-measure factors like parental influences – to be included in the X vector for the value-added estimates to be unbiased. As suggested by eq. (4), it is sufficient for the X vector to be rich enough that the unobserved factors do not vary much across teachers after controlling for X. But again, it is not clear whether this is likely.

however, they are really tracking & matching but their sorting practices based on observables and unobservables exactly offset each other. This is highly unlikely, as it would require π_1 and π_2 be of the opposite sign, or $\kappa < 0$ (observables and unobservables being negatively correlated). In our most prominent example, when X is lagged scores and Z is parental involvement, we naturally expect a strongly positive κ. Also, it is hard to think of any reasonable classroom assignment policies, where more assertive parents would want to put their kids into a classroom of low-achievers, which would result in π_1 and π_2 being of the opposite sign. Thus, I will assume away the possibility of $\pi_1 = -\kappa\pi_2 \neq 0$ for the rest of the paper. I thereby make schools that appear randomly assigning based on observables to constitute a benchmark case, where value-added scores are unbiased for true teacher quality.

On the contrary, at tracking & matching schools, where $\pi_1 + \kappa\pi_2 \neq 0$, the most likely scenario is $\pi_1, \pi_2 > 0$, yielding a nonzero bias in value-added scores.[11] For these schools, as well, it is theoretically possible that teacher effect estimates obtained from the partially controlled specification are unbiased but only if they only sort based on observables, and not on unobservables not predicted by observables, $\pi_1 \neq 0$, $\pi_2 = 0$. The following pieces of evidence suggest that this is unlikely.

Figure 1 plots estimates for μ obtained from a VAM with no controls at all and from a VAM with X controls (as in eq. (3)), separately for random and tracking & matching schools.[12] It shows that the two sets of estimates are nearly identical for random schools, suggesting no bias from observables in these schools (evidence for $\pi_1 + \kappa\pi_2 = 0$ in these schools). However, they are quite different for tracking & matching schools, with the scores from the uncontrolled specification suggesting positive sorting on observables (evidence for $\pi_1 + \kappa\pi_2 > 0$), that is, controlling for observables is important in tracking & matching schools. Therefore, we have a good reason to think that in these schools controlling for unobservables would also be important, were it possible (Altonji et al. [2005]; Oster [2014]). Table 7, discussed in more detail later,

[11] The magnitude depends on the controls included in the value-added specification, but generally it is an increasing function of $\beta, \rho, \pi_1, \pi_2, \kappa$ and the between-teacher variance of X and $\tilde{Z}$. In the partially controlled version, the magnitude of the bias depends on the variance of $\tilde{Z}$, π_2 and ρ. If any of these is small, the bias may be small as well. See the formal derivations in Appendix A.1.

[12] Data source and estimation details are discussed below, in Sections 3, 4 and 5.1.

supports this argument by showing that schools that track and match based on either lagged math or reading test scores are also more likely to track and match in other dimensions, and also that schools that are random based on either lagged math or reading scores are also more likely to be random on these other dimensions.

So the basis of my identification procedure is the following. As I can't identify sorting on unobservables, I detect sorting on observables. Then I argue that if there is no sorting on observables, it is very unlikely that there is sorting on unobservables. By contrast, when there is sorting on observables we need to at least worry that there may also be sorting on unobservables.

The following table shows the discussed cases of parameter values, with the more likely ones in bold underlined. Under these highlighted scenarios, we can identify the bias at tracking & matching schools by contrasting them with random schools. I will pursue on this strategy.

		Bias	
		Zero	Non-zero
School type	Random	$\underline{\pi_1 = \mathbf{0},\ \pi_2 = \mathbf{0}}$	$\pi_1 = -\kappa\pi_2 \neq 0$
	Tracking & Matching	$\pi_1 \neq 0,\ \pi_2 = 0$	$\underline{\pi_2 \neq \mathbf{0},\ \pi_2 \neq {}^{-\pi_1}/_{\kappa}}$

The strategy to identify the magnitude of bias in value-added scores is as follows.[13] Based on the argument above, the bias in randomly assigning schools is assumed to be 0, thus I assume that teacher value-added scores in randomly assigning schools has the following structure:

$$\bar{\tilde{\nu}} = \mu + \tilde{\theta} + \bar{\tilde{\varepsilon}},$$

where $\bar{\tilde{\nu}}$ is student residuals averaged to the teacher-year level, μ denotes true teacher value-added, $\tilde{\theta}$ is the classroom-level shock seen above, while $\bar{\tilde{\varepsilon}}$ is the average within-classroom noise. In contrast, estimates in tracking & matching schools take the form

$$\bar{\tilde{\nu}} = \mu + b + \tilde{\theta} + \bar{\tilde{\varepsilon}},$$

[13] See a more formal description in Section 5.2 below.

where b stands for the potential bias in true teacher value-added.[14] This structure implies several testable moment conditions, which identify the variance of the bias term and its covariance with the true effects. First, the difference of value-added score variances in the two types of schools pins down $Var(b) + 2Cov(\mu, b)$. Second, value-added autocovariances of teachers who stay at randomly assigning schools or move between two random schools in consecutive years identifies the variance of true teacher effects, $Var(\mu)$. Third, value-added autocovariances of teachers who move between a random and a tracking & matching school between consecutive years yields an estimate for $Var(\mu)+Cov(\mu, b)$. The difference between the last two terms, therefore, pins down $Cov(\mu, b)$, which in turn, together with $Var(b) + 2Cov(\mu, b)$ from the first observation identify $Var(b)$. After classifying schools into the two categories and estimating teacher value-added scores, I apply a minimum distance estimation strategy to complete the analysis.

3 Data

I use an administrative dataset of public K-12 schools provided by the North Carolina Education Research Data Center. These data contain the End-of-Grade test scores of 3rd, 4th and 5th graders from 1995 to 2011, among other variables. Students are linked to teachers who proctor the exam, however, for elementary school students these are most likely to be their classroom teacher. I estimate value-added models on a sample of teachers from a "base sample" of students for each test subject, reading and math. For a student-year observation to be included in the base sample for the given subject, the student has to have valid test scores in the given subject in the given year, lagged test scores in both subjects, and a valid teacher for the given subject.[15] A valid teacher is who has a class listed in the Student Activity Report in the given year, for the given grade level in the given subject. Self-contained classes are the dominant form of

[14] The variance of true teacher effects is assumed to be equal in the two types of schools, while I allow for different variances in classroom level shocks and noise terms.

[15] I also exclude students whose gender and race deviates from the mode in on third of the years available and students with certain disabilities. These disabilities are students with traumatic brain injuries, the mentally handicapped, the autistic, the deaf/blind, the visually impaired, the orthopedically impaired or the multi-handicapped. I include in my sample children who are behaviorally-emotionally handicapped, hearing impaired, speech-language impaired, specific learning disabled and other exceptional children.

instruction at elementary schools, and in the vast majority of cases, a valid math teacher is also a valid reading teacher. Nevertheless, for 2-3% of the student-year observations, the teacher is valid for only one of the subjects, which may be explained by earlier specialization in some schools, or coding errors. For most of my analysis, I measure teacher value-added in math, so I define a classroom as a school-grade-year-math teacher cell.

Summary statistics are shown in Table (1). 4th grade has the most observations due to missing test scores for 5th grade in 1996 and missing lagged scores for 3rd grade in a number of years. Lagged test scores in 3rd grade are actually not end-of-2nd grade scores but pretest scores from the beginning of 3rd grade. Such a test was not administered in every year.

The only control variables that are consistently available in every year are race and gender, frequency of homework and hours spent watching TV/using other electronics. In addition, I also use the following variables in years when they are available: parental education (available: 1995-2006), free/reduced price lunch (FRL) eligibility (1999-2006), exceptionality (1995-2009), and days absent (1995-2000). I use each of these as categorical variables with each value dummied out, including an indicator if they are missing.

Finally, an important feature of this dataset that I will exploit is that teachers can be followed as they move between schools, as long as they stay within the North Carolina public schools system.

4 Classroom Assignment Policies

In this section I describe the methodology I use to identify schools that practice random vs. nonrandom classroom assignment based on observable characteristics of students, and I discuss the results of this classification. I distinguish between two types of nonrandom assignment practices: tracking and matching. By tracking, I refer to what Rothstein [2009] describes as *dynamic tracking*: students each year are assigned to classes, based on newly observed attributes, such as test scores from the previous year.[16] More formally, this is equivalent to

[16]In this section, I describe nonrandom classroom assignment practices based on lagged test achievement. In Sections 4.1.2 and 4.2.1, I will investigate tracking and matching practices based on other observable student

testing whether a student's current classroom "predicts" their predetermined characteristics.

As opposed to tracking, which may only imply a contemporaneous correlation between student observables and the assigned teacher's quality, I define matching for detecting a potentially intertemporal correlation. Matching implies that a teacher each year is assigned to classrooms of students with similar observable characteristics. More formally, under matching, the identity of the current teacher "predicts" the average lagged achievement in her newly assigned classroom. To see why it is important to distinguish tracking from matching, consider the following example.

Let us take one grade level at a school in which each year there are two classrooms, taught by teachers j and k. If this school is tracking its students, in each year one of the classrooms will be filled up with the high achievers (based on last year's test scores), while the other with low achievers. If over time, both teachers j and k get assigned to the high achievers, the confounding effects of the assignments may "average out" and value-added estimates for the two teachers may be bias-free. However, if each year teacher j gets the high achievers and this fact also masks sorting based on unobservables, teacher effect estimates may be seriously biased. By the matching test, I want to see how prevalent this consistent assignment is across the years.

Two important thing have to be noted. First, from the example above it may appear, falsely, that tracking may not cause any problems in value-added estimation. Tracking is a prerequisite for matching, and as such deserve quantifying on its own right. Also, by identifying tracking as well, I will be able to more clearly distinguish between schools with both tracking and matching practices (to be referred to as "nonrandom" assigners) and those who engage in *neither* tracking or matching ("random" assigners).

Second, what follows is a purely positive analysis intended to assess the bias in teacher effect estimates in observational data. It is not suitable for evaluating whether tracking or matching

characteristics. I find that parental education may have a large influence in classroom assignments that is independent of lagged test scores. Nevertheless, when I quantify tracking and matching using predicted scores (an *index* of all predetermined characteristics from the student level value-added model), I end up with essentially the same pattern of nonrandom classroom practices as when I use only lagged scores.

are harmful or beneficial for students. In fact, matching teachers to certain types of classrooms may be an efficient allocation mechanism if it is based on comparative advantages, rather than seniority or nepotism.

4.1 Student Tracking

I quantify tracking based on the following test. I first run a regression of lagged scores on school-grade-year cells and classroom dummies in each school:

$$L\left(y_{it}\right) = \alpha_{sgt} + \kappa_c + u_{it}, \tag{7}$$

where $L(\cdot)$ is the lag operator, y_{it} is student i's test score in year t and $S(i,t) = s$, $G(i,t) = g$ and $C(i,t) = c$ are school, grade and classroom assignments.[17] The null hypothesis of *no tracking* is equivalent to testing that all κ_c are jointly 0 in each school,

$$H_0 : \ \kappa_c = 0 \ \forall c.$$

In other words, this is an F-test based on the ratio of the between-classroom and within-classroom variances with the appropriate degrees-of-freedom.[18] This test yields a p-value for each school. In Figure 2, Panels A and B, I plot the histogram of these p-values, in 5% bins.[19]

Figure 2 shows a striking spike in the first bin. If we adopt the convention of rejecting the null hypothesis of *no tracking* for p-values smaller or equal to 5%, we conclude that over 50% of schools engage in tracking.[20] The extreme left skew of the distribution provides clear evidence

[17]Classrooms in practice are defined as school-grade-year-teacher cells, so they are nested within school-grade-year cells.

[18]The numerator degrees-of-freedom is the overall number of classrooms in the school minus the number of school-grade-year cells, while the denominator degrees-of-freedom is the overall number of students in the school minus the overall number of classrooms. Alternatively, one could apply a Chi-squared test for the null that the classroom means within each school are equal. This test would be robust to heteroskedastic errors at the student level. However, this is a less powerful test if one wants to examine trends in tracking over time and across grade levels.

[19]Besides Clotfelter et al. [2006], Lefgren [2004] also performs a similar test, although he uses its output for a different purpose: He takes the R-squared of regression (7) as a continuous measure of tracking to identify student peer effects.

[20]If the null hypothesis of *no tracking* were true, the histogram would be a set of bars with equal, 0.05 height.

of tracking.

A similar histogram is shown in Figure (3), Panels A and B, separately school size.[21] Looking at the graphs, it appears that large schools are more likely to engage in tracking. However, that is not necessarily the case. The histogram bars for small schools drop off more gradually, which indicates that the test for small schools may not have as much power. When quantifying the prevalence of tracking, one has to take into account such power issues. I will do that in Section 4.1.1.

Note that a spike in the right-most bin would suggest that some schools intentionally *balance* the composition of their classrooms by lagged test scores. This is also a form of nonrandom assignment, for which I find very little evidence here.

4.1.1 The Power of the Test

To see if this test is actually suitable for measuring tracking, I have to account for potential power-related issues. Therefore, I run a simulation on a synthetic dataset, with the same number of student, school, grade, year and classroom observations as in the real dataset. I construct a synthetic lagged test score, and calibrate the data generating process for this variable so that the variance of the classroom and the school-grade-year components are the same as estimates in the real dataset. These variance components are shown in Table 2. Next, I perform the above F-test in each school for the simulated data, and plot the histogram of the p-values for all the schools. The resulting graphs are shown in Figures 4.

The first plot at the top shows that the p-values from the baseline specification, where none of the schools are tracking, have a uniform distribution. In contrast, when I assume that all schools track, I correctly predict this for 90% of the sample - suggesting my test has nearly 90% of power. For large schools (Panel B), my estimated power is even higher, essentially 100%. That is, among schools with above median size that practice tracking (defined as having a p-value less than or equal to 0.05 in the above F-test), I correctly predict this with 100%

[21] I define large and small schools as follows. I take all students in the base sample who went to the given school across all years and grade levels. Large schools will be the ones who have above median number of students overall, while small schools are the below median ones.

probability. Among small schools, the test does a slightly worse job detecting tracking only 80% of the time.[22]

Returning to my results in Figure (3), my power tests suggest that nearly 60% of large schools track their students. Since I find that I can correctly detect tracking among small schools only 80% of the time, results in Figure (3) likely understate true tracking. Adjusting for this, I find that $^{0.5}/_{0.8} = 0.625$ of the small schools are tracking, which is remarkably similar to what we found for large schools. Overall, I conclude that about 60% of *all* schools engage in tracking.[23]

4.1.2 Tracking on Other Observables

So far I have been focusing on tracking based on lagged test scores, however, it is interesting to look at tracking on other predetermined variables. Here, I compute the same tracking statistic using female/male, white/non-white, college educated parents/lower educated parents and free- or reduced price lunch (FRL) eligible/not eligible dummies on the left handside of eq. (7). Again, I classify schools as tracking on the given variable if the p-value obtained from the F-test of within- and between-classroom variances is less then or equal to 0.05. All these statistics are computed using only school-grade-year cells where there are sufficient number of students from both groups.[24] Results are shown in Appendix Figure A.1. One striking result is that the gender composition of classrooms is extremely balanced, as shown by the huge excess mass in the right tail of the p-value distribution in Panel A. Panel B also shows evidence of about 40% of the schools balancing on race. I also find a small portion of schools that appear

[22]The results of the simulation can also be summarized in a table that shows the percent correctly predicted from the test, when tracking is defined as having a p-value of less than or equal to 0.05. Table 3 shows these results.

I also performed the simulation exercise with a *mix* of tracking and non-tracking schools, which reveals both type-I and type-II errors related to the test. I classified 60% of the schools as tracking, as suggested by the tracking test results in Section 4.1. The misclassification table for this specification is shown in Appendix Table A.1. The table reveals that the share of correctly specified schools is around 93%, another piece evidence suggesting that this test of tracking is very powerful. Furthermore, the test appears rather conservative, as there are somewhat more tracking schools that are not detected to be tracking than non-tracking schools misclassified as tracking.

[23]Recall that tracking is formally defined as having a p-value less than or equal to 0.05 in the test above.

[24]I require at least 15% white and nonwhite, 15% FRL eligible and non-eligible students, and at least 10% with and without college educated parents.

to be tracking on race. This may be the result of either: intentional segregation, or, and more likely, tracking based on race actually due to tracking based on lagged scores. I will check this below. In Panel C, I find a large number of schools tracking on parental education, even more than the number of those that track on lagged scores. This may signal that more involved parents have stronger preferences about classrooms within schools. Finally, about 40 of schools track on FRL eligibility, although some of this may also be due to tracking on lagged scores.

Correlations of my measures of tracking on different observable characteristics are displayed in Table A.2. Unsurprisingly, tracking on lagged math scores is highly correlated with tracking on lagged reading scores. This means that in elementary schools, classrooms either do not specialize in math or reading, or children's knowledge in the two subjects are not easily differentiated by the test itself. Correlation between tracking on lagged scores and other socio-economic background variables is smaller, but still positive.

To check how much of the tracking based on these demographic and socio-economic characteristics is independent of tracking on lagged scores, I run the same regressions as in eq. (7) with the above four student characteristics on the left handside, adding lagged math and reading scores as controls. I perform the same F-test for the joint significance of classroom dummies and again obtained p-values for each school. These "conditional" p-values are plotted in histograms separately for schools that do track on either math or reading lagged scores, and for schools that do not. Figure 5 shows the results. Strong balancing on gender is robust to controlling for lagged scores, as is balancing on race, especially among schools that do not track on lagged scores. Notice that the few schools that appeared earlier to be tracking on race stop doing so once we control for tracking on lagged scores. Nevertheless, the vast majority of tracking on parental education or FRL status holds up, even after taking into account tracking based on lagged scores. Moreover, tracking on parental education and FRL status is more prevalent among schools that track on lagged scores as well. This suggests that classroom assignment may be based on factors unobservable to the researcher as well.

4.2 Teacher Matching

Having identified the schools that are tracking their students, a prerequisite for matching teachers to classes, let us see next how many of these schools engage in matching. Recall that intuitively, schools that match tend to assign similar classrooms of students to the same teacher over time. To quantify the prevalence of this, I perform the following test of matching using average lagged math score at the classroom level.

I collapse the student-level data to the classroom level and run a regression of average lagged scores in the classroom on school-grade-year cell and teacher dummies in each school:

$$\overline{L\left(y_c\right)} = \alpha_{sgt} + \mu_j + u_c, \tag{8}$$

where $L(\cdot)$ is the lag operator, so $\overline{L\left(y_c\right)}$ is the average lagged test score in classroom c. $S(c,t) = s$, $G(c,t) = g$ denote the school and grade in which the classroom is observed, and $J(c,t) = j$ is the teacher assigned to classroom c.[25] The regressions are weighted by class size. The null hypothesis of *no matching* is equivalent to testing that all μ_j are jointly 0 in each school,

$$H_0 : \ \mu_j = 0 \ \forall j.$$

which is an F-test on the ratio of between-teacher and within-teacher variances with the appropriate degrees-of-freedom. [26]This test yields a p-value for each school, and I plot the histogram of these p-values in 5% bins in Figure (6) Panels A and B.

These graphs have a similar interpretation as the tracking results in Figures 2 and 3. Thus, if we defined matching as having a p-value of less than or equal to 0.05 in the above test, then we would classify 40% of all schools matching. Recall, however, that schools can logically only engage in matching if they also engage in tracking. Therefore, it is more useful to look at the

[25]Note that since a teacher is observed across years (and potentially schools and grade levels), j is not nested within school-grade-year cells.

[26]The numerator degrees-of-freedom is the overall number of teachers in the school minus the number of school-grade-year cells , while the denominator degrees-of-freedom is the overall number of classrooms in the school minus the overall number of teachers.

prevalence of matching among tracking schools. This is shown in Figure 7, Panels A and B, which reveals that almost 50% of tracking schools also engage in matching teachers persistently to certain types of classrooms. If schools did not match according to my definition, then the graph would yield a uniform distribution. The left-skew suggests clear evidence of matching. However, two other features of the graph indicate that this test may not be very powerful. First, the height of the bars are gradually decreasing beyond the first bin. Second, I also find a non-negligible fraction of non-tracking schools matching. I return to these issues and the power of the test in Section 4.3 below. For now, note that if we look at matching among large schools only, where we would expect the test to have more power, we see that 60% of tracking schools also engage in matching (see Figure 7 Panels C and D).

Interestingly, there appears to be no schools that intentionally try to *balance* teacher assignments over time. In these figures, schools that balance teacher assignments would be indicated by an excess mass at the right tail of the distribution.

4.2.1 Matching on Other Observables

It may be of interest to look at whether schools persistently match teachers using other observable characteristics as well. Therefore, I run the school-level matching regressions, eq. (8), with dependent variables percent female, percent white, percent with college educated parents and percent free- or reduced price lunch eligible students in the classroom. Again, these regressions are run only using school-grade-year cells in which there are a sufficient proportion of student from each group. Appendix Figure A.2 shows the histogram of p-values for the F-test with "no matching" as the null hypothesis, among schools that engage in tracking based on the given variable.[27] As we saw in Section 4.1.2, there are very few schools that track on gender or race. Therefore, the matching histograms for these two variables are meaningless, I just report them for the sake of completeness. There is some evidence of matching on FRL status, however,

[27]The same graphs for schools that do not engage in tracking based on the given variable are hard to interpret, as tracking is a "prerequisite" for matching. Nevertheless, these graphs are shown in Appendix Figure A.3. Panels B-D show some evidence for matching in schools that track on lagged scores but so small that it can be due to misclassification.

this is mostly prevalent in schools that track based on lagged test scores. In contrast, matching certain teachers to classroom with high vs. low proportion of kids with college educated parents seems a common practice, even in schools that do not track students by lagged scores. This means that my final categorization may understate the number of tracking & matching schools. On the one hand, that would suggest that nonrandom assignment potentially causing biases in teacher value-added scores are even more prevalent than measured here. On the other hand, even in that case, the approach I take will underestimate biases in teacher value-added score that are due to nonrandom sorting.

When I compute the matching statistics using these demographic and socio-economics characteristics additionally controlling for lagged math and reading scores, the plots of p-values (Figure 8[28]) reveal that the small evidence of matching on FRL composition we saw above is washed away.[29] That is, teachers are matched on average previous achievement, which confounds with FRL composition. On the other hand, matching on parental education is robust to controlling for lagged scores and appears even among schools that randomly assign their students based on lagged scores. This pattern is consistent with anecdotical evidence that highly educated parents try to get their kids into the classrooms with the same teachers year after year.

4.3 Simulation: Joint Power of The Tracking and The Matching Tests

We saw some signs in the previous subsection that the matching test may have low power. To check this, I run a simulation, similar to that described in Section 4.1.1, but this time for the two tests, tracking and matching, jointly. As before, I use the number of student, classroom, and school-grade-year observations from the real data, and the DGP is generated so that the variance components of the synthetic lagged scores reflect their estimates in the

[28]The same graphs for schools that do not engage in tracking based on the given variable are shown in Appendix Figure A.4, and reassuringly, do not show any evidence of matching.

[29]Similarly to above, the sample of schools that track on gender or race is very small, so Panels A-B are only reported for the sake of completeness.

real data. In the simulation, I generate 3 types of schools: (i) schools that assign students and teachers randomly to classrooms ("random"); (ii) those that track students but do not match teachers to classrooms ("tracking only"); and (iii) those that perform both tracking and matching ("tracking & matching"). The proportion of each type reflects the results of the classifications in sections 4.1 and 4.2: Students at a random 60% of schools have both classroom and school-grade-year components in their scores (tracking schools), while the remaining 40% have only a school-grade-year component (randomly assigning schools). Students at a random half of tracking schools were also given a time-invariant teacher component in their scores (tracking & matching schools). Table 4 summarizes the features of this data generating process for lagged scores.

I run the tracking test described above on the simulated scores and then collapse the data to the classroom level and run the matching test. As a result, each school obtains a p-value for tracking and a p-value for matching. I classify schools as tracking (matching) if their tracking (matching) p-value is less than or equal to 0.05, and I summarize the results in Table 5, a three-way misclassification table that shows the percent correctly predicted.

Summing the values along the diagonal, the table shows that 74.7% of schools are correctly classified. Misclassification arises as a serious issue only in one case: schools that are generated to be both trackers and matchers. About 60% of these schools are correctly classified as tracking but not as matching as well. It is interesting to note that schools that only reject matching and not tracking (however, tracking is logically a presumption for matching) but not tracking appear to come from to the truly randomly assigning group.

Table 6 shows the empirical distribution of schools in the dataset, where I have classified schools into 3 types:

1. *Random*ly assigning schools: schools for which both the tracking and the matching tests (in both math and reading) failed to reject the null at the 0.05 level (upper left cell in Table 6).[30]

[30]Technically, schools that reject the "no matching" test only (upper right cell in Table 6) also belong in this

2. *Tracking-only* schools: schools for which the tracking test is rejected at the 0.05 level, but the matching test fails to reject the null (lower left cell in Table 6). I will not use this set of schools in the analysis below.

3. *Tracking & matching* (or "nonrandom") schools: schools for which I reject the null at the 0.05 level for both the tracking and matching tests in either math or reading (lower right cell in Table 6). This may be the group for whom teacher value-added estimates may be the most severely biased.

In the analysis I present below, I focus only on two types of schools, random and tracking & matching (sometimes also referred to as nonrandom). As argued in Section 2, the teacher effects in the random group are likely to be measured nearly bias free, while those in the nonrandom group are most likely to be contaminated by a bias term.

These two groups are quite robust even if we measure tracking & matching based on (i) a weighted average of lagged math and reading scores (where the weights come from the coefficients on the lagged score variables in a student-level value-added model), or (ii) predicted scores from all observable variables in a student-level value-added model. As Table 7 shows, there are essentially no random assignment schools on lagged scores that appear tracking & matching on these other two variables, while there are almost no tracking & matching schools on lagged scores that appear random on these variables. Classification based on a weighted average of observables other than lagged scores (including parental education for instance) performs somewhat worse, however, as Table 7 suggests, classification on lagged scores is rather conservative: there are more random schools on lagged scores that appear tracking & matching on other observables, and relatively few who are tracking & matching based on lagged scores but random on other observables. This again supports that nonrandom assignment may be more prevalent than quantified here and so my approach may just yield a lower bound on biases in teacher value-added scores introduced by nonrandom sorting.

group, as the simulation shows they are misclassifications in the truly randomly assigning group. I decide to not use this group of schools in the rest of the analysis, but the results are robust to their inclusion as random assigners.

5 Structure of Teacher Value-Added

Now that I have classified randomly and nonrandomly assigning schools, I turn to how such a classification can be used to identify the magnitude of the bias in teacher effect estimates, and its correlation with true teacher effects. I begin by describing how I obtained value-added scores for each teacher linked to the students in my base sample.

5.1 Student-level VAM specification

I obtain teacher value-added scores using the "average residual" method: After estimating a (student-level) value-added model, I average the student-level residuals at the classroom level.

The student-level VAM specification I use is common in the literature:[31]

$$y_{it} = \gamma_g \left(y_{ig-1t-1} \right) + X_{it}^{'} \beta + \delta_{gt} + \nu_{it}. \tag{9}$$

In all specifications the dependent variable is the math test score of student i in year t, and $S\left(i,t\right) = s$, $G\left(i,t\right) = g$ and $J\left(i,t\right) = j$ are the student's school, grade and teacher assignments in the given year. Lagged math and reading scores enter the right handside as a flexible, grade-specific, 3rd order polynomial, $\gamma_g \left(y_{ig-1t-1} \right)$. δ_{gt} is a set of grade×year effects to control for changes in the tests and other grade- and year-specific shocks. Control variables, X_{it}, are gender and race of a child, 4 categories of parental education, eligibility for free or reduced price lunch, exceptionality status in 8 categories, time spent on TV/video game, frequency of doing homework and days of absence. Only gender, race, homework and TV/other electronics usage are available for all years in the base sample, so I include a missing indicator for each of the rest of the controls when they are unavailable.[32] I estimate (9) by OLS, with standard errors clustered at the school level. I carry out 3 additional robustness checks. First, I add "leave-out" classroom average peer characteristics and "leave-classmates-out" school-grade-year average

[31] For a full review of different specifications that are commonly estimated in the literature, see Guarino et al. [2012].

[32] See more details about the availability of each control variable in the Data section.

peer characteristics to eq. (9).[33] Second, I augment the specification by school-grade-year effects, and last, I add teacher effects (omitting school-grade-year effects). In this last within-teacher specification, before proceeding, I add back the estimated fixed effects for teachers to the student-level residuals, similar to how Chetty et al. [2014a] and Bacher-Hicks et al. [2014] did.

After obtaining $\hat{\nu}_{it}$ from eq. (9), I compute (inefficient) teacher value-added scores by averaging student-level residuals for each teacher-year combination:[34]

$$\bar{\nu}_{jt} = \frac{1}{n_{jt}} \sum_{i \text{ s.t. } J(i,t)=j} \hat{\nu}_{it}.$$

There are a small number of teachers in my base sample who show up teaching multiple classes in the same year.[35] In their case, I take the precision weighted average of their classroom-level residuals, so that I end up with one observation per teacher-year. Table 8 contains the summary of implied teacher value-added estimates. These simple value-added scores, with no shrinkage, are appropriate to estimate the structure of value-added.

5.2 Identifying the bias and its relationship with the true teacher effect

Assume there are two types of schools: those in which classroom assignment practices are random ($P(s) = R$) and those that both track their students and match teachers persistently

[33] These average characteristics are lagged scores, percent of each minority, percent female, percent of parental education categories, percent free- and reduced price eligible, percent gifted and percent other type of exceptional. I take these averages at the classroom level, leaving out the student herself, and also average them over the whole school-grade-year cell, leaving out the student and her classmates.

[34] A common practice to compute teacher value-added scores is to combine multiple years of data and use shrinkage. (See, for instance, Jacob et al. [2010], Kane and Staiger [2008], Chetty et al. [2014a], Rothstein [2014], Bacher-Hicks et al. [2014]. However, for my purpose, simply averaging over student-level residuals in a single year fits best.)

[35] Teachers with more than one classroom assignment make up 1.12% of my base sample. I did not throw them out, because in theory it may be possible that in some schools teachers specialize in certain subjects even at these early grades, so these teachers may only teach one subject but multiple classes. Also, there may be schools with mixed grade-level classrooms. Leaving these teachers out from the analysis does not change any of the results.

to similar classrooms over time.[36] Again, I call this latter group nonrandom assigners ($P(s) = TM$). Residuals from the student-level VAM in eq. (9) can be decomposed as follows:

$$\nu_{it} = \mu_j + b_{jP(s)} + \theta_{jP(s)t} + \varepsilon_{it},$$

where ν_{it} is the residual from the student-level value-added model (VAM) for student i assigned to teacher $J(i,t) = j$ in year t at school $S(i,t) = s$ with classroom assignment policy $P(s)$. μ_j is the true teacher effect for teacher j, $b_{jP(s)}$ is the bias term in teacher j's effect if teacher j is at school type $P(s)$. θ_{jt} is a classroom-level shock, which is assumed to be $i.i.d.$ in both the teacher and time dimensions. This captures noise that is common to all students in the classroom (e.g. someone jackhammering outside when the kids are taking the test, or a teacher-year shock). ε_{it} is student-level $i.i.d.$ noise. Then average teacher-year residuals are

$$\bar{\nu}_{jt} = \mu_j + b_{jP(s)} + \theta_{jP(s)t} + \bar{\varepsilon}_{jt}.$$

I assume there is no bias in randomly assigning schools. However, there is bias potentially in tracking & matching schools. That is,

$$b_{jR} \equiv 0,$$

$$b_{jTM} \equiv b_j \neq 0,\ \sigma_b^2 > 0,\ \sigma_{\mu b} \neq 0.$$

Using this structure, we can compute second moments (variances and autocovariances) of teacher value-added estimates for five different groups of teachers:

1. Variance of value-added scores for all teachers in randomly assigning schools:

$$Var\left(\bar{\nu}_{jt} \mid P(s) = R\right) = \sigma_\mu^2 + \sigma_{\theta_R}^2 + \frac{\sigma_{\varepsilon_R}^2}{n_{jt}}.$$

In practice, I will take the adjusted variance of $\bar{\nu}_{jt}$ for the within-classroom variance, $\frac{\sigma_{\varepsilon_R}^2}{n_{jt}}$

[36] I discuss how I measure these two types in Section 4.

:

$$\begin{aligned} Var\left(\text{adj}\bar{\nu}_{jt} \mid P(s)=R\right) &= E\left[\left(\bar{\nu}_{jt} - \frac{\sum_{j\in R}\bar{\nu}_{jt}}{\#\text{ of R teachers}}\right)^2 - \frac{\sigma^2_{\varepsilon_R}}{n_{jt}}\right] = \\ &= \sigma^2_{\mu} + \sigma^2_{\theta_R}, \end{aligned}$$

with appropriate degree-of-freedom adjustments.

2. Variance of value-added scores for all teachers in tracking & matching schools. Similarly adjusting for the within-classroom variance,

$$\begin{aligned} Var\left(\text{adj}\bar{\nu}_{jt} \mid P(s)=TM\right) &= E\left[\left(\bar{\nu}_{jt} - \frac{\sum_{j\in TM}\bar{\nu}_{jt}}{\#\text{ of TM teachers}}\right)^2 - \frac{\sigma^2_{\varepsilon_{TM}}}{n_{jt}}\right] = \\ &= \sigma^2_{\mu} + \sigma^2_{b} + 2\sigma_{\mu b} + \sigma^2_{\theta_{TM}}. \end{aligned}$$

3. Autocovariance of stayers in random schools and movers between two random schools:[37]

$$Cov\left(\bar{\nu}_{jt}, \bar{\nu}_{jt-1} \mid P\left(S(j,t)\right) = P\left(S(j,t-1)\right) = R\right) = \sigma^2_{\mu}.$$

4. Autocovariance of stayers in tracking & matching schools:

$$Cov\left(\bar{\nu}_{jt}, \bar{\nu}_{jt-1} \mid S(j,t) = S(j,t-1) \text{ and } P\left(S(j,t)\right) = TM\right) = \sigma^2_{\mu} + \sigma^2_{b} + 2\sigma_{\mu b}.$$

5. Autocovariance of movers from randomly assigning schools to nonrandomly assigning ones and vice versa:

$$Cov\left(\bar{\nu}_{jt}, \bar{\nu}_{jt-1} \mid P\left(S(j,t)\right) \neq P\left(S(j,t-1)\right)\right) = \sigma^2_{\mu} + \sigma_{\mu b}.$$

[37] As ε's are assumed to be i.i.d., within-classroom variance terms do not appear in the covariances, therefore it is not necessary to adjust for them.

I back out the empirical counterpart of these moments in the data and perform a minimum distance estimation to obtain estimates for the underlying parameters,

$$\Theta = \left(\sigma^2_\mu, \sigma^2_b, \sigma_{\mu b}, \sigma^2_{\theta_R}, \sigma^2_{\theta_{TM}}\right).$$

The theoretical moments alongside with their empirical counterparts are listed in Table 9). They can be turned into a just identified system of 5 moment conditions with 5 parameters to estimate. Then the criterion function to be minimized is

$$\min_{\Theta} \left[\hat{m} - f\left(\Theta\right)\right]' W \left[\hat{m} - f\left(\Theta\right)\right],$$

where $\hat{m}$ is the vector of empirical moments and $f\left(\Theta\right)$ is a vector of theoretical moments. W is a weighting matrix. In the just-identified case it is simply the identity matrix. To obtain standard errors, I compute the variance-covariance matrix from 1000, 90% block bootstrapped samples of empirical moments.

5.3 Results

The results of the minimum distance estimation outlined in the previous subsection are shown in Table 10. I only report results from the plain OLS student-VAM specification.[38]

I find that the standard deviation of the true teacher effects is 0.175, which is in the high end of the range in the previous literature (around 0.1-0.2 standard deviations [SDs]).[39] Classroom-level shocks (0.13-0.14 SDs) are substantial relative to true teacher effects, and are remarkably similar across the two school types.

One of the main results of this paper is the estimate of the variance of the potential bias

[38]Results for the other three student-level VAM specifications are qualitatively the same and quantitatively very similar. Also, minimum distance results are only reported for the case when schools are classified based on lagged math and reading scores, however, these are almost the same, both qualitatively and quantitatively, as results for the case when school are classified based on predicted scores from all observable variables.

[39]In the other three specifications, I find 0.09-0.21 SDs. The smallest SD is in the within school-grade-year specification, while the largest is in the within-teacher one (where teacher fixed effect estimates are added back to student-level residuals).

in teacher value-added scores, which turns out to be quantitatively important: it is almost a third of the variance of true effects, with 0.093 SDs, although not statistically significant.

I find a substantially negative correlation between true teacher effects and the bias in value-added scores, around -0.3. This is the first estimate of this parameter. It suggests that previous literature evaluating the bias in teacher effects due to nonrandom classroom assignment but assuming zero correlation may have been misleading. (See more on this point in the next subsection.) The negative sign of the correlation between teacher effects and the bias indicates a negative sorting of students to teachers based on characteristics that are unobservable given observed student covariates.[40] In other words, good teachers tend to get students whose potential learning gains are lower, and bad teachers get assigned to students whose potential gains are higher. This finding is in contrast with the positive sorting process based on observables that we saw in Figure 1, implying that students who are predicted to score higher based on their observable characteristics are assigned to good teachers, but in turn will acquire lower gains, plausibly due to mean reversion. This mechanism eventually hurt their teachers' measured effectiveness. In fact this implies that bad teachers are helped by value-added score measurements, while good teachers are punished. Due to the direction of this mismeasurement, even the rank ordering of teacher may be affected.

In the methods of moment estimation, I treat teachers who stay at randomly assigning schools in two consecutive years and those who move between two randomly assigning school as one group, assigning them the same theoretical autocovariance parameter. Also, I treat teachers who move between a randomly assigning school and a non-randomly assigning school in either direction to be identical. These assumptions are testable. Therefore, I run regressions of current value-added score on lagged value-added scores, fully interacted with a categorical variable that can take 5 values: (i) the teacher is a stayer at a random school, (ii) the teacher is stayer at a

[40] This is an interesting finding on its own, and points into the direction that teacher assignments may not be efficient within schools.

Using the notation from Section 2, this result means that $Cov\left(\mu, \tilde{Z}\right) < 0$, implying $\pi_2 - \kappa\pi_1 < 0$ in tracking & matching schools. If assignment practices in random schools are qualitatively similar in random schools as well, just less strong, this condition precludes the knife-edge case of $\pi_1 + \kappa\pi_2 = 0$, when my identification of bias would not work.

tracking & matching school, (iii) the teacher moves between two random schools, (iv) the teacher moves from a random school to a tracking & matching one, and (v) the teacher moves from a tracking & matching school to random one. The coefficients on the interaction terms in this regression inform us about the differences in the reliability of year-to-year value-added scores for each stayer/mover type of teacher. The stayer/mover types for whom the reliabilities are significantly different should be parametrized differently in the minimum distance estimation. Results are shown in Table 11. The first column shows OLS estimates, while the second column adds current school effects interacted with year effects. The reference group is teachers who stay in the same randomly assigning schools in two consecutive years.

The reliability of the value-added scores is low (around 0.53); however, this is only due to the fact that I am only using one year of data to predict value-added in the subsequent year. There are only economically small and borderline significant main effects, however, these are only due to between-school sorting of teachers and are absorbed by (current) school effects. The coefficient on lagged value-added score gives the 1-year reliability of value-added scores among teachers who stay in random schools (the reference group), while the coefficients on the interaction terms are the differences in the reliabilities for each group of teachers, relative to the reference group. There is one coefficient that directly contrasts with the model I applied in the minimum distance estimation: The reliability of the random-to-random movers is significantly lower than that of the random school stayers. The reliability of groups 4 and 5 also seem somewhat different, however, this difference is not significant. To see whether my model of value-added structure is robust to these differences, I estimate an overidentified extension separating the group of random school stayers from random-to-random movers and random-to-nonrandom movers from nonrandom-to-random movers, thereby allowing for two extra moment conditions. The moment conditions are listed in Table 12. The estimation results are shown in Table 13.

As the model is now overidentified, I use three different weighting matrices: an identity matrix, a diagonal matrix with the number of observations in the diagonal and the inverse variance-covariance matrix of the moments (the optimal weighting matrix). Table 13 shows that the results are relatively robust to the choice of the weighting matrix. The specification

with the identity weighting matrix yields somewhat lower bias variance and lower correlation between the true effects and the bias, but this specification does not take into account the fact that the moments for the movers are much less precisely estimated than those for the stayers, due to the smaller sample sizes. Overall, the results are similar to what I found in Table 10: there may be substantial biases in estimated teacher value-added, however, those biases are mostly offset by their negative correlation with the true effects. This negative correlation is what made the bias undetectable using the methods employed earlier in the literature.[41]

5.4 Discussion

I have provided a new estimate of the bias in teacher value-added scores, and for the first time in the literature, assessed the correlation between this bias and the true teacher effects. Both turn out to be quantitatively important. In particular, the negative sign of the correlation suggests that the methods used to estimate value-added may also *rank* teachers incorrectly. These results are somewhat at odds with recent studies evaluating the performance of teacher value-added scores using teacher switching quasi-experiments (Kane and Staiger [2008], Kane et al. [2013];Chetty et al. [2014a], Bacher-Hicks et al. [2014]). This conflict is due to the misleading interpretation of results in these earlier studies.

To assess the reliability of teacher value-added scores, researchers have focused on the coefficient in a regression of true teacher effects on value-added scores,

$$\frac{Cov\left(\mu_j, \hat{\mu}_j\right)}{Var\left(\hat{\mu}_j\right)}.$$

The teacher switching studies estimate a feasible version of this coefficient assuming random classroom assignment: the regression coefficient of a change in student achievement at a school-grade on a change in teacher value-added due to teacher movements at the same school-grade. They find a coefficient very close to 1 and interpret it as evidence for no bias in value-added

[41] The overidentification test does not reject (p-value= 0.51) the model.

scores.[42]

However, this interpretation is only valid under the implicit assumption that true teacher effects are uncorrelated with the potential bias in the measured value-added. Formally, assuming $\hat{\mu}_j = \mu_j + b_j$ yields

$$\frac{Cov\left(\mu_j, \hat{\mu}_j\right)}{Var\left(\hat{\mu}_j\right)} = \frac{Var\left(\mu_j\right) + Cov\left(\mu_j, b_j\right)}{Var\left(\mu_j\right) + Var\left(b_j\right) + 2Cov\left(\mu_j, b_j\right)}, \tag{10}$$

and by assuming $Cov\left(\mu_j, b_j\right) = 0$, this expression simplifies to

$$\frac{Cov\left(\mu_j, \hat{\mu}_j\right)}{Var\left(\hat{\mu}_j\right)} = \frac{Var\left(\mu_j\right)}{Var\left(\mu_j\right) + Var\left(b_j\right)}, \tag{11}$$

which in turn implies $Var\left(b_j\right) \approx 0$ if $\frac{Cov(\mu_j, \hat{\mu}_j)}{Var(\hat{\mu}_j)} = 1$.[43]

In this paper, I took a different approach to estimate $\frac{Cov(\mu_j, \hat{\mu}_j)}{Var(\hat{\mu}_j)}$ by recovering the underlying parameters in (10). Most importantly, my approach directly estimates $Cov\left(\mu_j, b_j\right)$, instead of implicitly assuming it away. For teachers whose value-added scores are measured in random schools, $Var\left(b_j\right) = Cov\left(\mu_j, b_j\right) = 0$, therefore my results suggest a coefficient of 1. However, in tracking & matching schools the coefficient turns out to be around 0.87, with $Cov\left(\mu_j, b_j\right) < 0$ and $Var\left(b_j\right) \neq 0$. This number is within the confidence intervals of the true experiments analyzed in Kane and Staiger [2008] and Kane et al. [2013], who find forecast coefficients of 0.8-1 and are almost never able to reject that they are 1. My estimate is also not significantly different from 1, but at the same time, it's magnitude is remarkably similar to what Rothstein [2014] found, when revisiting the validity of the quasi-experiment in the Chetty et al. [2014a] study.[44] These findings suggest that interpreting the reliability of value-added scores according to (11) may be misleadingly simplistic.[45] According to the traditional interpretations of the forecast

[42] In Chetty et al. [2014a]'s terminology, this amounts to *forecast unbiasedness*.

[43] Chetty et al. [2014a] calls this *teacher-level unbiasedness*. They discuss that this is equivalent to forecast unbiasedness except in the knife-edge case where $Var\left(b_j\right) + Cov\left(\mu_j, b_j\right) = 0$ and $Var\left(b_j\right) \neq 0$.

[44] He shows that teachers (and their students) who are left out of the Chetty et al. [2014a] analysis are not random - these are teacher who are in their sample for at most the two years over which they analyze the grade- or school switches. After correcting for this nonrandom selection, he finds a coefficient around 0.8-0.9, statistically significant from 1.

[45] In fact, when I estimate the more restrictive structure of value-added assuming away $Cov\left(\mu_j, b_j\right)$, I get back the result $\frac{Cov(\mu_j, \hat{\mu}_j)}{Var(\hat{\mu}_j)} = \frac{Var(\mu_j)}{Var(\mu_j)+Var(b_j)} = 1$ implying $Var\left(b_j\right) \approx 0$. This model has a worse fit, however,

coefficient, a value of 0.87 would imply a $\frac{Var(b_j)}{Var(\mu_j)}$ ratio of 0.15, which is about half of what I find by estimating the richer model that allows for a nonzero covariance between μ_j and b_j. That interpretation masks a potentially much larger bias in value-added scores. The contrast between that and my findings reveals that even if teacher value-added scores are unbiased predictors of the student achievement (and yield a forecast coefficient indistinguishable from 1), they may be so for the wrong reason: they may contain a large teacher-level bias, part of which is offset by its negative covariance with true teacher effects. This point relates to Chetty et al. [2014a]'s discussion that there is one knife-edge case in which value-added scores are best predictors but biased for the causal effect for teachers, and that is when $Var(b_j) + Cov(\mu_j, b_j) = 0$ and $Var(b_j) \neq 0$.[46] My estimates suggests that the truth may not be far off from this knife-edge case, which always lies in the confidence interval. (See the last line in Tables 10 and 13, which show that $Var(b_j) + Cov(\mu_j, b_j) = 0$ cannot be rejected at conventional significance levels.) As a consequence, value-added scores may be a useful statistical tool to predict student achievement on average, however, they may not be accurate enough to be used for individual teacher evaluation or for personnel decisions in a fair way.

In my setup, I am able to empirically distinguishing between $Var(\mu_j)$ and $Cov(\mu_j, b_j)$ using teacher movements between randomly assigning schools and tracking & matching schools. These movements, however, are relative rare, as shown by the number of observations in Table (12). Previous studies such as Chetty et al. [2014a] might have failed to detect this potentially large bias in value-added scores because they identified it predominantly from teacher switches within school types and from grade switches within schools. My results say that within-school, the reliability of value-added estimates may be close to 1, however, they may be quite a bit lower (around 0.87) when a teachers moves from a random school to a tracking & matching one, or vice versa. This suggests that the reliability of teacher value-added scores highly depend on the classroom assignment policy that schools or school systems employ. Previously, heterogeneity in assignment practices were ignored, and researchers have estimated an *average* of the forecast

than the richer structure allowing for $Cov(\mu_j, b_j) \neq 0$.

[46] This occurs when the forecast coefficient, $\frac{Cov(\mu_j, \hat{\mu}_j)}{Var(\hat{\mu}_j)} = \frac{Var(\mu_j)+Cov(\mu_j,b_j)}{Var(\mu_j)+Var(b_j)+2Cov(\mu_j,b_j)} = 1$ but $Var(b_j) \neq 0$, implying the condition $Var(b_j) + Cov(\mu_j, b_j) = 0$ with $Var(b_j) \neq 0$.

coefficient across the school types. My results suggest that about half of the clearly classifiable schools are random and the other half are tracking & matching, and that about 60% of teachers in these schools work in tracking & matching schools and 40% work in random schools. By back-of-the-envelope calculations, if we estimate an average forecast coefficient across these teachers, we would get $0.4 \times 1 + 0.6 \times 0.87 = 0.92$, which would least to underestimating biases in nonrandom assignment schools. Furthermore, if we interpret this coefficient under the implicit assumption that the biases are uncorrelated with the true effects (as described in more detail above), we would get a bias-true variance ratio of 0.09, about a third of what I found in this paper.

6 Conclusion

This paper contributes to a highly controversial literature on the accuracy of teacher value-added measures. Recently researchers have argued that such measures are reliable enough to be used for high stakes personnel decisions, such as promotions of performance-based rewards. In this paper, I estimated the bias in teacher value-added measures in North Carolina elementary schools, employing a new strategy that uses teachers who move between schools that assign their classes randomly and schools that track students and match teachers to classrooms with similar kids year after year. I classified schools into these two categories using variance decomposition techniques. About a third of all schools were found to engage in both tracking and matching based on lagged test scores, while another third were classified as random assigners. Beyond previous achievement, I found evidence that parental education may be another, independent and strong factor influencing classroom assignment. This is in line with anecdotical evidence that more involved parents actively fight for a preferred teacher.

Based on this classification, I argued that teacher effect estimates in randomly assigning schools are likely to be nearly bias free, while in nonrandomly assigning schools, they may be contaminated by bias. A minimum distance framework was applied to recover the underlying parameters of the structure of teacher value-added. Importantly, I provided the first estimates

for the correlation between the potential bias and true teacher effects. According to my findings, the variance of the bias in estimated teacher value-added may be as large as 30% of the variance of true teacher effects. However, its large negative correlation with the true teacher effect mitigates the forecast bias in value-added scores. These results confirm Rothstein [2014]'s findings that the reliability of value-added scores may not be unbiased predictors of student achievement and adds two important details to our knowledge about value-added estimation. First, that forecast unbiasedness is context-dependent: Value-added scores may be very reliable in schools or school systems that randomly assign their students to teachers, but considerably less so in tracking & matching schools. Second, even if they are unbiased predictors of student achievement, they may not be unbiased estimators for the causal effect of teachers. Therefore, their application for personnel decisions is unjustified. Teacher value-added may be a powerful *statistical* tool within districts made up of schools with similar assignment practices but one has to apply caution when using them for policies that evaluate *individual* teachers, such as tenure or performance pay.

The findings in this paper also shed some light on the within-school classroom assignment mechanism in schools with nonrandom assignment practices: On average students with observables that predict high test scores get assigned to good teachers. Potentially due to mean reversion, these students will then be likely to have lower gains in the following year. This is an interesting pattern on its own that calls for more research.

References

D. Aaronson, L. Barrow, and W. Sander. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 2007.

J. G. Altonji, T. E. Elder, and C. R. Taber. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, 2005.

J. Alzen and B. Domingue. A characterization of sorting and implications for value-added estimates. Working Paper, 2013.

A. Bacher-Hicks, T. J. Kane, and D. O. Staiger. Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles. NBER Working Paper No. 20657, 2014.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, 2014a.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 2014b.

C. T. Clotfelter, H. F. Ladd, and J. L. Vigdor. Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources*, 2006.

S. G. Dieterle, C. M. Guarino, M. M. Reckase, and J. M. Wooldridge. How do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-Added. Working paper, The Education Policy Center at Michigan State University, 2013.

E. Duflo, P. Dupas, and M. Kremer. Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 2011.

C. M. Guarino, M. M. Reckase, and J. M. Wooldridge. Can Value-Added Measures of Teacher

Performance Be Trusted? Working paper, The Education Policy Center at Michigan State University, 2012.

E. A. Hanushek. Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data. *American Economic Review*, pages 280–288, 1971.

D. N. Harris and D. R. Sass. Teacher Training, Teacher Quality, and Student Achievement. Working paper, CALDER Urban Institute, 2007.

B. A. Jacob and L. Lefgren. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, pages 101–136, January 2008.

B. A. Jacob, L. Lefgren, and D. Sims. The Persistence of Teacher-Induced Learning Gains. *Journal of Human Resources*, pages 915–943, 2010.

D. Kalogrides, S. Loeb, and T. Béteille. Systematic Sorting: Teacher Characteristics and Class Assignments. *Sociology of Education*, 2013.

T. J. Kane and D. O. Staiger. Estimating Teacher Impacts On Student Achievement: An Experimental Evaluation. NBER Working Paper No. 14607, 2008.

T. J. Kane, D. F. McCaffrey, T. Miller, and D. O. Staiger. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Technical report, Bill & Melinda Gates Foundation, 2013.

L. Lefgren. Educational Peer Effects and the Chicago Public Schools. *Journal of Urban Economics*, pages 169–191, 2004.

B. Nye, S. Konstantopoulos, and L. Hedges. How large are teacher effects? *Educational Evaluation and Policy Analysis*, 2004.

E. Oster. Unobservable Selection and Coecient Stability: Theory and Evidence. 2014.

S. G. Rivkin, E. A. Hanushek, and J. F. Kain. Teachers, Schools, and Academic Achievement. *Econometrica*, pages 417–458, 2005.

J. E. Rockoff. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review: Papers and Proceedings*, pages 247–252, 2004.

J. Rothstein. Student Sorting And Bias In Value-Added Estimation: Selection On Observables And Unobservables. *Education Finance and Policy*, pages 537–571, 2009.

J. Rothstein. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, pages 175–214, 2010.

J. Rothstein. Revisiting the Impacts of Teachers. Working Paper, 2014.

J. Rothstein and W. J. Mathis. Review of Two Culminating Reports from the MET Project. Technical report, National Education Policy Center, 2013.

Tables

Table 1: Summary Statistics of the Student-Level Data

		N	mean	sd	min	max	years available
Test scores	mean(math)	2,842,551	0.077	0.971	-4.958	3.037	1995-2011
	mean(reading)	2,840,369	0.068	0.968	-4.575	3.149	
Race	white	2,842,551	0.618				1995-2011
	black		0.269				
	hispanic		0.059				
	asian		0.018				
	other		0.036				
Gender	female	2,842,551	0.501				1995-2011
Parental education	parent <HS	1,991,512	0.097				
	parent=HS		0.424				
	parent some coll		0.211				
	parent coll or more		0.269				
FRL status	free lunch elig	1,416,426	0.346				1999-2006
	reduced price lunch elig		0.093				
Exceptionality	gifted	2,542,416	0.122				1995-2009
	other exceptional (7 cats.)		0.084				
TV/other electronics	None	2,826,174	0.070				1995-2011
	1-5 hours (3 cats.)		0.836				
	6+ hours a school day		0.094				
Homework	0-1 hours a week (2 cats.)	2,829,539	0.344				1995-2011
	1-3 hours a week		0.338				
	3 to 10+ hours a week		0.309				
	has hw but does not do		0.009				
Days absent	0-7 days	918,765	0.719				1995-2000
	8-20 days (2 cats.)		0.262				
	21+ days		0.019				
Grade level	grade=3	719,558					1997-2005, 2007, 2009
	grade=4	1,100,351					1995-2011
	grade=5	1,022,642					1995, 1997-2011
Observations	num student-years	2,842,551					
	num students	1,487,863					
	num teachers	38,472					
	num classrooms	154,547					
	num school-grade-years	50,478					
	num schools	1,543					

Note: Student-level base sample. Years used differ by grade level due to availability of current and lagged scores. Math and reading scores are standardized to have mean 0 and standard deviation 1 by each grade and year, for all non-missing observations. The base sample is still somewhat selective, as it contains only student-year observations, where teachers to math classroom was successfully matched, the student had non-missing lagged test scores and were not classified exceptional because of traumatic brain injuries, mental handicap, autism, being deaf/blind, visually impairment, orthopedic impairment or being multi-handicapped. Other exceptional status includes children who are behaviorally-emotionally handicapped, hearing impaired, specific learning disabled, speech-language impaired and other health impaired.

Table 2: Simulation of Tracking: Variance of Test Score Components in DGP

proportions	type	Variance Components		
		classroom	school-grade-year	noise
0%	non-tracking	-	$\mathcal{N}(0, 0.35^2)$	$\mathcal{N}(0, 1 - 0.35^2)$
100%	tracking	$\mathcal{N}(0, 0.25^2)$	$\mathcal{N}(0, 0.35^2)$	$\mathcal{N}(0, 1 - 0.35^2 - 0.25^2)$

Note: Variance components used to simulate the power of the tracking test. The data generating process for a synthetic, student-level lagged test score was specified using random draws from the corresponding distributions in the table, according to

$$L(y_{it}) = \alpha_{sgt} + \kappa_c + u_{it},$$

where $\alpha_{sgt} \sim \mathcal{N}(0, 0.35^2)$ is the school-grade-year component, $\kappa_c \sim \mathcal{N}(0, 0.25^2)$ is the classroom component and $u_{it} \sim \mathcal{N}(0, 1 - 0.35^2 - 0.25^2)$ is noise. See more details in Section 4.1.1.

Table 3: Simulation of Tracking: Percent Correctly Predicted and Misclassification

# schools=1,415	don't reject	reject	Total
small schools	0.199	0.801	1.000 (0.483)
large schools	0.002	0.998	1.000 (0.517)
Total	0.097	0.903	1.000

Note: Share of correctly and misclassified schools in the simulation of the tracking test, using the number of student, classroom, school and school-grade-year observations from the real data. Statistics are shown by school size, as the test is likely to perform worse for small schools. Large (small) schools are defined to have above (below) median number of students from the base sample across all years. The more powerful the test is, the higher the rejection rate is relative to row totals. This test has essentially 100% power for large schools and about 80% power for small schools. See more details in Section 4.1.1.

Table 4: Simulation of Tracking and Matching: Variance of Test Score Components in DGP

proportions	type	Variance Components			
		classroom	teacher	school-grade-year	noise
40%	random	-	-	$\mathcal{N}(0, 0.35^2)$	$\mathcal{N}(0, 1 - 0.35^2)$
30%	tracking but not matching	$\mathcal{N}(0, 0.25^2)$	-	$\mathcal{N}(0, 0.35^2)$	$\mathcal{N}(0, 1 - 0.35^2 - 0.25^2)$
30%	tracking and matching	$\mathcal{N}(0, 0.18^2)$	$\mathcal{N}(0, 0.15^2)$	$\mathcal{N}(0, 0.35^2)$	$\mathcal{N}(0, 1 - 0.35^2 - 0.18^2 - 0.15^2)$

Note: Variance components used to simulate the power of the joint tracking-matching test. 40% of schools are prescribed to be randomly assigning, 30% to track students only, while 30% to track students and match teachers persistently to classrooms. The data generating process for a synthetic, student-level lagged test score was specified using random draws from the corresponding distributions in the table, according to

$$L(y_{it}) = \alpha_{sgt} + \kappa_c + \mu_j + u_{it}.$$

The tracking test was then performed and afterwards the synthetic data was collapsed to the classroom level to perform the matching test. See more details in Section 4.3.

Table 5: Simulation of Tracking and Matching: Percent Correctly Predicted and Misclassification

	reject neither	reject "no tracking"	reject "no matching"	reject both	Total
random	0.897	0.051	0.051	0.000	1.000 (0.409)
tracking	0.058	0.884	0.000	0.058	1.000 (0.310)
matching	0.000	0.000	0.000	0.000	0.000
tracking-matching	0.050	0.569	0.004	0.377	1.000 (0.281)
Total	0.399	0.455	0.022	0.124	1.000

Note: Share of correctly and misclassified schools in the joint simulation of the tracking and matching test, using the number of student, classroom, school and school-grade-year observations from the real data. 40% of schools were prescribed to be random assigners, 30% to only track students, and 30% to both track students and persistently match teachers to classrooms. No schools were prescribed to match only. The more powerful the test is, the higher the diagonal elements should be relative to row sums. Total percent correctly predicted is 74.7%. The tracking part has almost 90% power, however, the matching has only about 40%. The joint test is rather conservative though, as there are more false non-rejections then false rejections. See more details in Section 4.3.

Table 6: Tracking and Matching: Empirical Distribution and Classification

# schools (% schools)	non-matching	matching	Total
non-tracking	414 (29.3)	133 (9.4)	547 (38.7)
tracking	433 (30.6)	435 (30.7)	868 (61.3)
Total	847 (59.9)	568 (40.1)	1,415 (100.0)

Note: Based on the performed variance decomposition tests, this table shows the empirical distribution of schools according to their classroom assignment practices. Further on, I will classify schools that are not tracking students (40.2%) as "random assigners", while those that both track students and persistently match teachers to classrooms, as "nonrandom assigners" (or "tracking & matching"; 31.6%). I will not use schools that are tracking (28.2%) in the rest of the analysis. See more details in Section 4.3.

Table 7: Different Measures of Tracking & Matching

VARIABLES	(1) R on weighted lagged scores	(2) TM on weighted lagged scores	(3) R on pred score	(4) TM on pred score	(5) R on other observables	(6) TM on other observables
R on lagged scores	0.959	0.000	0.925	0.000	0.452	0.171
	(0.010)	(0.000)	(0.013)	(0.000)	(0.024)	(0.019)
TM on lagged scores	0.007	0.885	0.007	0.892	0.057	0.685
	(0.004)	(0.015)	(0.004)	(0.015)	(0.011)	(0.022)
Observations	849	849	849	849	849	849
R-squared	0.952	0.885	0.918	0.892	0.405	0.586

Note: Robust standard errors in parentheses. Linear probability models, school-level base sample. Dependent variables: indicators of random (in odd columns) or tracking & matching (in even columns) on (1) weighted average of lagged reading and math scores, with weights being the coefficients of lagged scores in the student-level VAM; (2) predicted score, from all observables (see details in Section 5.1); and (3) weighted average of observables other than lagged scores, with weights being the coefficients of these observables in the student-level VAM. Independent variable: indicator for tracking & matching on lagged scores. The sample consists of schools that can be classified into the 2 groups based on lagged scores (independent variable), have non-missing dependent variable but may not be clearly classifiable to R or TM based but may not be clearly classifiable based on the dependent variable. See details on classification in Section 4.

Table 8: Summary of Teacher Value-Added Estimates

	OLS	OLS with peer characteristics	School-grade-year effects	Teacher effects
Unadjusted SD	0.245	0.242	0.191	0.285
Adjusted SD	0.223	0.221	0.166	0.266
N	152,826	151,227	151,227	151,227

Note: Column titles denote the student-level VAM specification. Teacher value-added scores are measured by average student-level residuals at the teacher-year level. In the very few cases where a teacher is assigned to multiple classrooms a year, precision weighted averages of the classroom averages are reported. See more details in Section 5.1. Adjusted standard deviations are computed by netting out the within-classroom component (sampling variation) from the unadjusted value-added score variance.

Table 9: Structure of Teacher Value-Added

Moment Type	School/Movement type	Theoretical Moment	Empirical Moment	Observations
Adjusted variances	R	$\sigma_\mu^2 + \sigma_{\theta_R}^2$	0.0494	38,105
	TM	$\sigma_\mu^2 + \sigma_b^2 + 2\sigma_{\mu b} + \sigma_{\theta_{TM}}^2$	0.0478	56,736
Autocovariances	stayers in R, movers R-to-R	σ_μ^2	0.0307	23,963
	stayers in TM	$\sigma_\mu^2 + \sigma_b^2 + 2\sigma_{\mu b}$	0.0297	35,208
	movers R-to-TM and TM-to-R	$\sigma_\mu^2 + \sigma_{\mu b}$	0.0258	809

Note: Theoretical moment conditions derived from value-added structure, $\bar{\nu}_{jt} = \mu_j + b_{jP(s)} + \theta_{jP(s)t} + \bar{\varepsilon}_{jP(s)t}$, where μ_j is a teacher-level component, $b_{jP(s)}$ is a school-type specific bias component ($b_{jR} \equiv 0$, $b_{jTM} \equiv b_j$ with $\sigma_b^2 > 0$), $\theta_{jP(s)t}$ is a school-type specific classroom-level shock, and $\bar{\varepsilon}_{jP(s)t}$ is the classroom-average of school-type specific, student-level noise. Variances are adjusted by within-classroom component, $Var\left(\bar{\varepsilon}_{jP(s)t}\right) \equiv \frac{\sigma^2_{\varepsilon_{P(s)}}}{n_{jt}}$. School types are randomly assigning school (R), and tracking & matching school (TM, also called as nonrandomly assigning).

Table 10: Structure of Teacher Value-Added - Estimation Results for 5 Moment Conditions

Parameters	Estimates
Var of true teacher effect, $\hat{\sigma}^2_{\mu}$	0.0307 (0.0005)
Var of bias, $\hat{\sigma}^2_{b}$	0.0087 (0.0060)
Cov of true effect and bias, $\hat{\sigma}_{\mu b}$	-0.0049 (0.0030)
Var of classroom shock, $\hat{\sigma}^2_{\theta_R}$	0.0187 (0.0007)
Var of classroom shock, $\hat{\sigma}^2_{\theta_{TM}}$	0.0181 (0.0005)
# of parameters	5
# of moments	5
Implied SD(true effect)	0.1752
Implied SD(bias)	0.0933
Implied SD(classroom shock$_R$)	0.1367
Implied SD(classroom shock$_{TM}$)	0.1345
Implied $Corr$(true effect, bias)	-0.2998
Implied forecast coefficient$_R$, $\left(\frac{\hat{\sigma}^2_{\mu}}{\hat{\sigma}^2_{\mu}}\right)$	1
Implied forecast coefficient$_{TM}$, $\left(\frac{\hat{\sigma}^2_{\mu}+\hat{\sigma}_{\mu b}}{\hat{\sigma}^2_{\mu}+2\hat{\sigma}_{\mu b}+\hat{\sigma}_{b^2}}\right)$	0.8704 (0.1012)
$\hat{\sigma}^2_{b}+\hat{\sigma}_{\mu b}$	0.0038 (0.0030)

Note: Minimum distance estimation results for the underlying parameters of the teacher value-added structure. Standard errors in parentheses. Standard errors for the implied parameters were computed using the delta method. R subscripts denote components in randomly assigning schools, while TM subscripts denote components in nonrandomly assigning (tracking & matching) schools. The bias term is allowed to be contained only in value-added estimates in nonrandom schools. True teacher effects are assumed to have the same distributional characteristics in both types of schools. Results are identical when using 3 different weighting matrices (identity, observation numbers in the diagonal or optimal) due to being just-identified. When $\hat{\sigma}^2_{b}+\hat{\sigma}_{\mu b}=0$ and $\hat{\sigma}^2_{b}\neq 0$, teacher value-added scores are unbiased predictors of student achievement but biased estimates for the causal effect of teachers.

Table 11: Reliabilities of Teacher Value-Added - 5 groups

VARIABLES	(1) value-added score	(2) value-added score
lagged value-added score (VAS; stayers in R schools)	0.530	0.536
	(0.010)	(0.011)
lagged VAS*group2: stayers in TM schools	0.004	0.006
	(0.012)	(0.014)
lagged VAS*group3: movers R-to-R schools	-0.122	-0.165
	(0.043)	(0.051)
lagged VAS* group4: movers R-to-TM schools	-0.064	-0.114
	(0.064)	(0.077)
lagged VAS* group5: movers TM-to-R schools	-0.131	-0.165
	(0.051)	(0.060)
group 2: stayers in TM schools	-0.006	0.000
	(0.003)	(0.014)
group 3: movers R-to-R schools	0.019	-0.009
	(0.014)	(0.015)
group 4: movers R-to-TM schools	-0.006	
	(0.011)	
group 5: movers TM-to-R schools	0.022	0.003
	(0.013)	(0.016)
Constant	0.014	0.010
	(0.002)	(0.008)
Observations	59,980	59,980
R-squared	0.283	0.515
school-year effects	no	yes

Note: Standard errors, clustered at the school level, in parentheses. Regressions weighted by average class size. Dependent variable: teacher value-added scores (average teacher-year residuals from student-level VAM). Regressions show the year-to-year reliability of teacher value-added scores for the different groups of teachers. Reference group: teachers staying in random (R) schools.

Table 12: Structure of Teacher Value-Added - Overidentified Version

Moment Type	School/Movement type	Theoretical Moment	Empirical Moment	Observations
Adjusted variances	R	$\sigma_\mu^2 + \sigma_{\theta_R}^2$	0.0494	38,105
	TM	$\sigma_\mu^2 + \sigma_b^2 + 2\sigma_{\mu b} + \sigma_{\theta_{TM}}^2$	0.0478	56,736
Stayers' autocovariances	R	σ_μ^2	0.0308	23,528
	TM	$\sigma_\mu^2 + \sigma_b^2 + 2\sigma_{\mu b}$	0.0297	35,208
Movers' autocovariances	R-to-R	σ_μ^2	0.0264	435
	R-to-TM	$\sigma_\mu^2 + \sigma_{\mu b}$	0.0255	403
	TM-to-R	$\sigma_\mu^2 + \sigma_{\mu b}$	0.0265	406

Note: Theoretical moment conditions derived from value-added structure, $\bar{\nu}_{jt} = \mu_j + b_{jP(s)} + \theta_{jP(s)t} + \bar{\varepsilon}_{jP(s)t}$, where μ_j is a teacher-level component, $b_{jP(s)}$ is a school-type specific bias component ($b_{jR} \equiv 0$, $b_{jTM} \equiv b_j$ with $\sigma_b^2 > 0$), $\theta_{jP(s)t}$ is a school-type specific classroom-level shock, and $\bar{\varepsilon}_{jP(s)t}$ is the classroom-average of school-type specific, student-level noise. Variances are adjusted by within-classroom component, $Var\left(\bar{\varepsilon}_{jP(s)t}\right) \equiv \frac{\sigma_{\varepsilon_{P(s)}}^2}{n_{jt}}$. School types are randomly assigning school (R), and tracking & matching school (TM, also called as nonrandomly assigning).

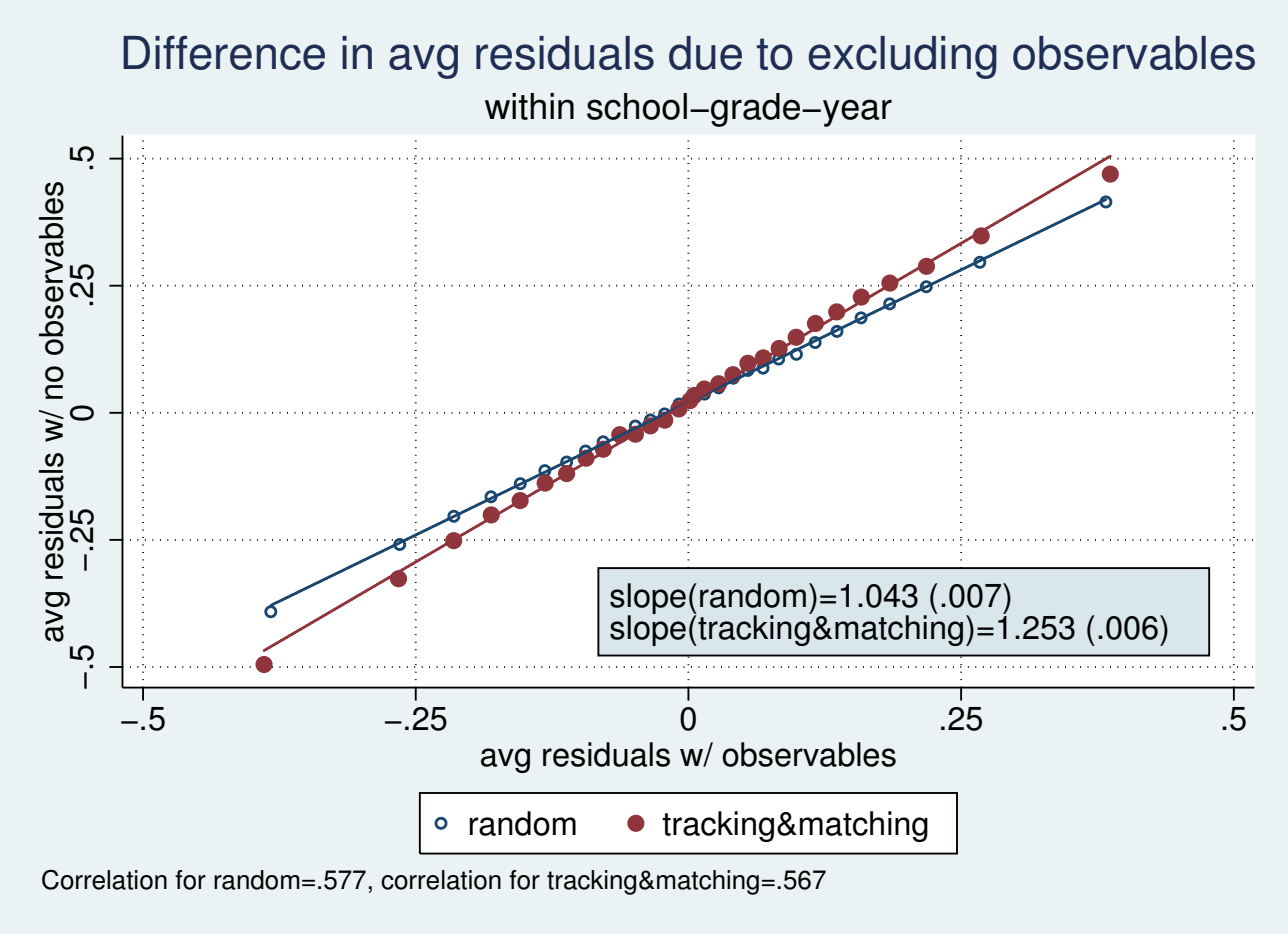
Table 13: Structure of Teacher Value-Added - Estimation Results with 7 Moment Conditions

Parameters	identity matrix	obs. numbers in the diagonal	optimal weighting matrix
Var of true teacher effect, $\hat{\sigma}^2_{\mu}$	0.0288 (0.0019)	0.0307 (0.0005)	0.0308 (0.0005)
Var of bias, $\hat{\sigma}^2_{b}$	0.0063 (0.0063)	0.0084 (0.0060)	0.0081 (0.0060)
Cov of true effect and bias, $\hat{\sigma}_{\mu b}$	-0.0026 (0.0035)	-0.0047 (0.0030)	-0.0046 (0.0030)
Var of classroom shock, $\hat{\sigma}^2_{\theta_R}$	0.0208 (0.0020)	0.0187 (0.0007)	0.0187 (0.0007)
Var of classroom shock, $\hat{\sigma}^2_{\theta_{TM}}$	0.0181 (0.0005)	0.0181 (0.0005)	0.0181 (0.0005)
SSE			1.3439
# of parameters	5	5	5
# of moments	7	7	7
Implied SD(true effect)	0.1691	0.1752	0.1755
Implied SD(bias)	0.0794	0.0917	0.0900
Implied SD(classroom shock$_R$)	0.1422	0.1367	0.1367
Implied SD(classroom shock$_{TM}$)	0.1345	0.1345	0.1345
Implied $Corr$(true effect, bias)	-0.1937	-0.2927	-0.2912
Implied forecast coefficient$_R$, $\left(\frac{\hat{\sigma}^2_{\mu}}{\hat{\sigma}^2_{\mu}}\right)$	1	1	1
Implied forecast coefficient$_{TM}$, $\left(\frac{\hat{\sigma}^2_{\mu}+\hat{\sigma}_{\mu b}}{\hat{\sigma}^2_{\mu}+2\hat{\sigma}_{\mu b}+\hat{\sigma}_{b^2}}\right)$	0.8751 (0.1017)	0.8752 (0.1017)	0.8811 (0.1009)
$\hat{\sigma}^2_{b}+\hat{\sigma}_{\mu b}$	0.0037 (0.0030)	0.0037 (0.0030)	0.0035 (0.0030)

Note: Minimum distance estimation results for the underlying parameters of the teacher value-added structure. Standard errors in parentheses. Standard errors for the implied parameters were computed using the delta method. R subscripts denote components in randomly assigning schools, while TM subscripts denote components in nonrandomly assigning (tracking & matching) schools. The bias term is allowed to be contained only in value-added estimates in nonrandom schools. True teacher effects are assumed to have the same distributional characteristics in both types of schools. When $\hat{\sigma}^2_{b}+\hat{\sigma}_{\mu b}=0$ and $\hat{\sigma}^2_{b}\neq 0$, teacher value-added scores are unbiased predictors of student achievement but biased estimates for the causal effect of teachers.

Figures

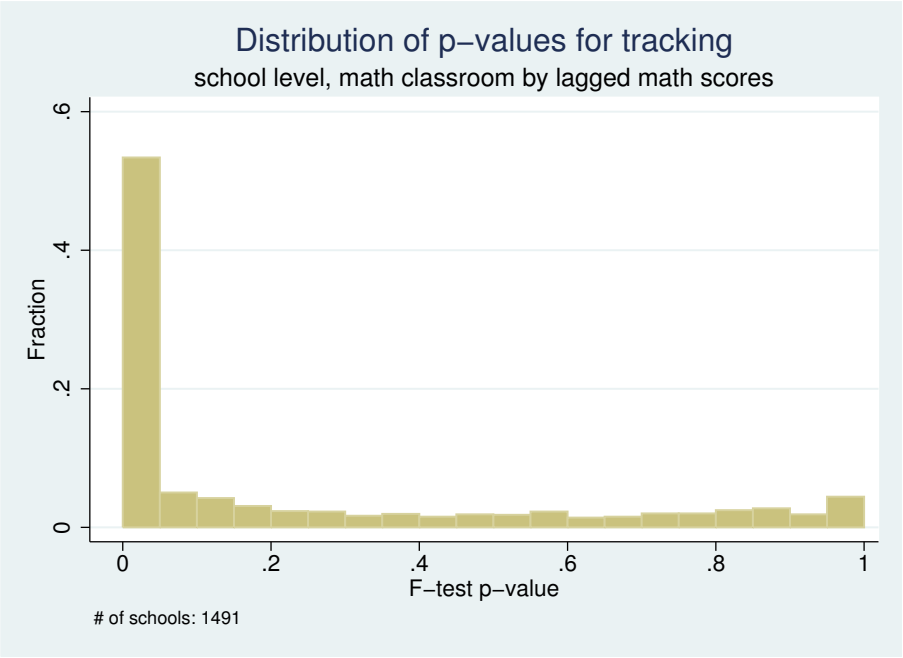
Figure 1: Unshrunken teacher value-added estimates with and without controlling for observables (random vs. tracking & matching schools)



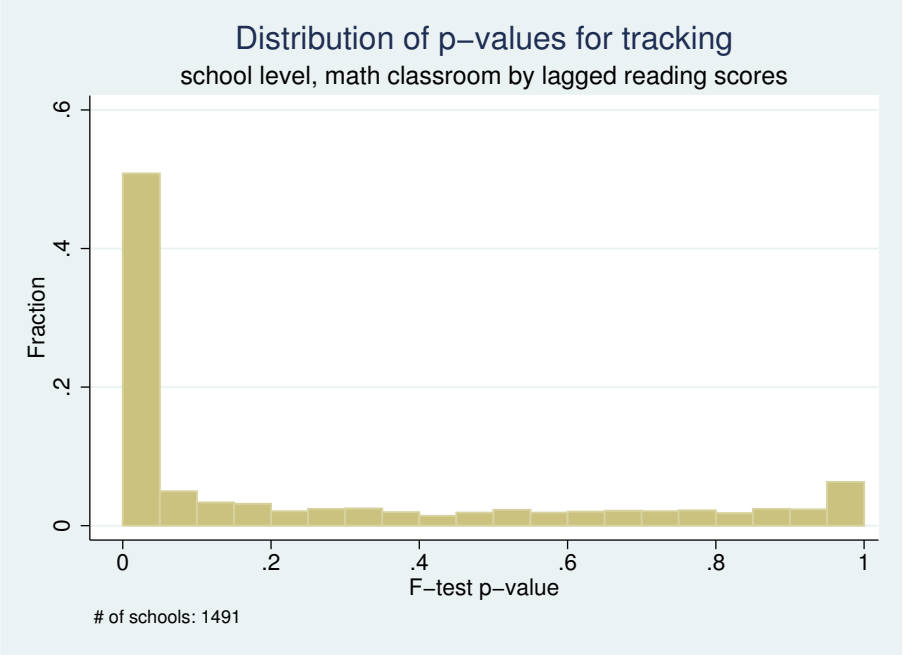
Note: Binned scatter plots of teacher value-added scores, defined as the teacher-year average of student-level residuals. Specification with no observables is simply teacher-demeaned current year achievement. Specification with observables are obtained from residuals from OLS student-VAM, eq. (9).

Figure 2: Histogram of p-values for no tracking

Panel A: By Lagged Math Scores



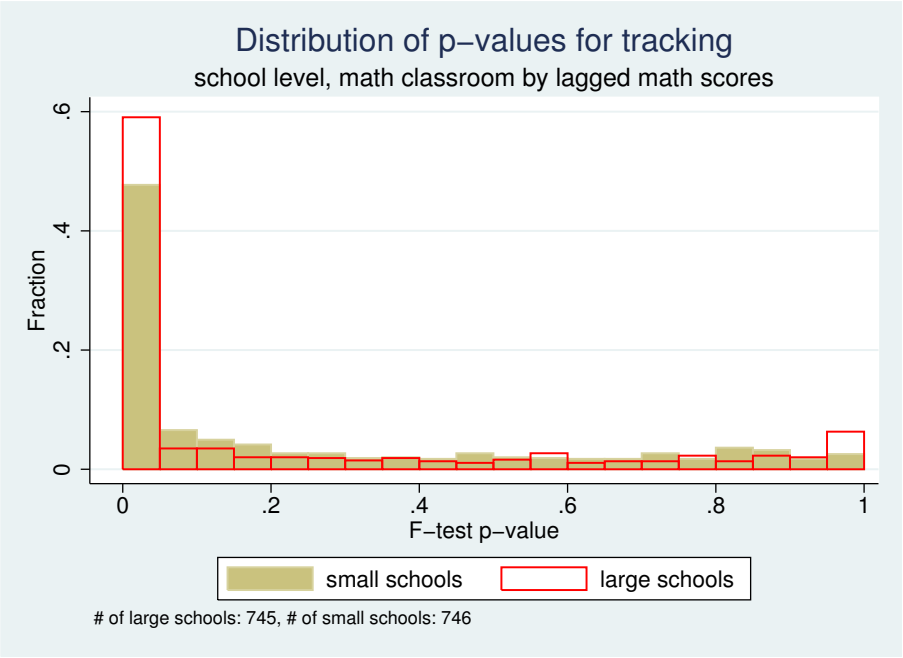
Panel B: By Lagged Reading Scores



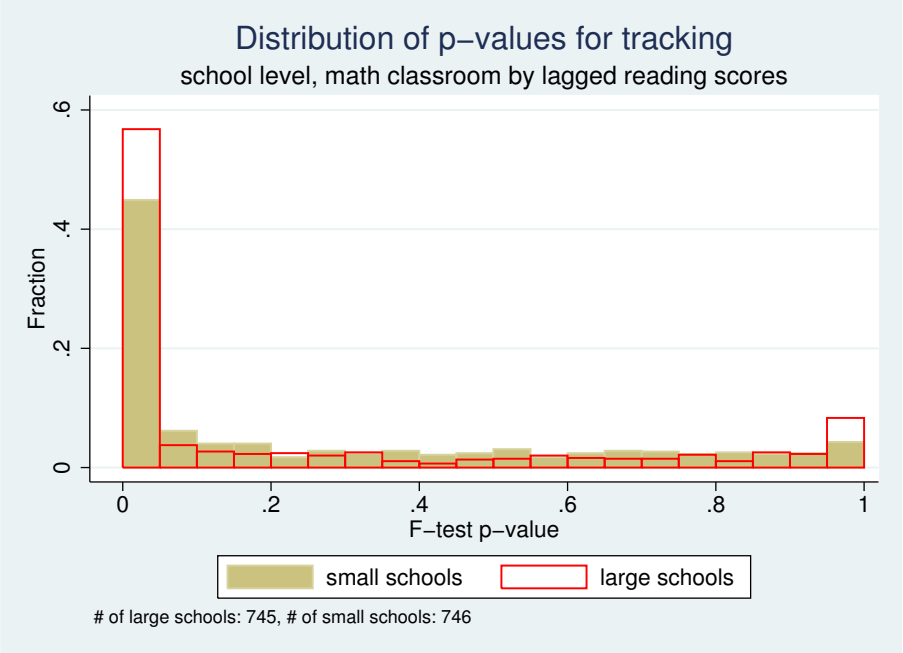
Note: Distributions of p-values are displayed from F-test on the joint significance of classroom effects within school-grade-year cells. Student-level regressions for each school, dependent variables: lagged math or reading scores. Null hypothesis: no tracking (current classroom assignment does not predict prior achievement). Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as tracking students. Schools in the last bin are the few ones who intentionally balance their classes based on prior achievement.

Figure 3: Histogram of p-values for no tracking by school size

Panel A: By Lagged Math Scores



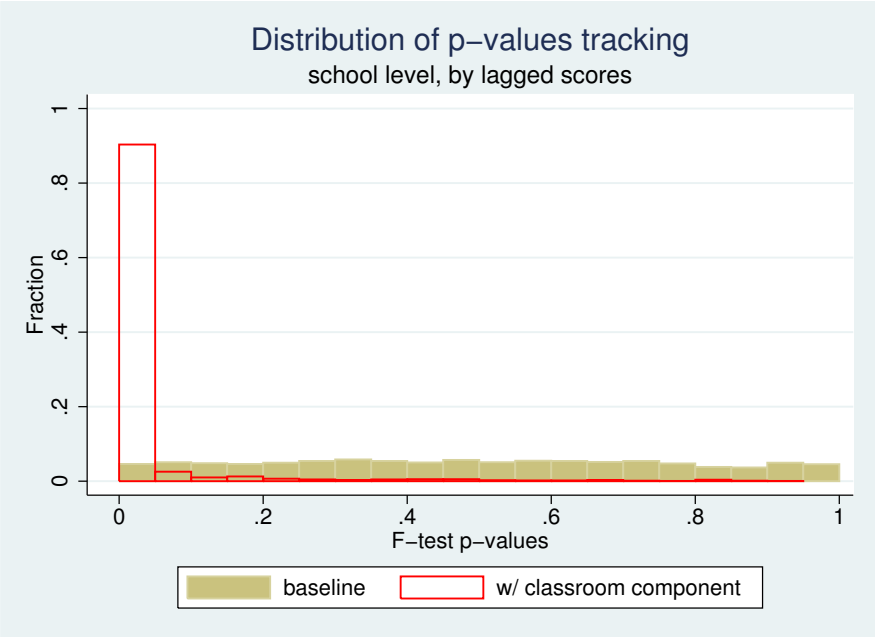
Panel B: By Lagged Reading Scores



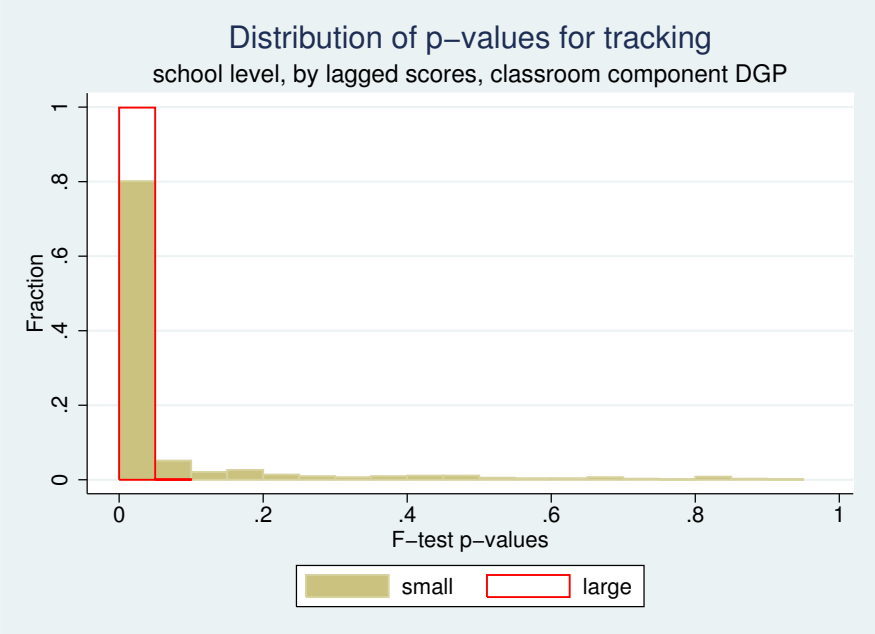
Note: Distributions of p-values are displayed from F-test on the joint significance of classroom effects within school-grade-year cells. Student-level regressions for each school, dependent variables: lagged math or reading scores. Null hypothesis: no tracking (current classroom assignment does not predict prior achievement). Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as tracking students. Schools in the last bin are the few ones who intentionally balance their classes based on prior achievement. Large (small) schools are defined to have above (below) median number of students from the base sample across all years.

Figure 4: Simulation: Histogram of p-values for no tracking

Panel A: All schools, tracking vs. non-tracking



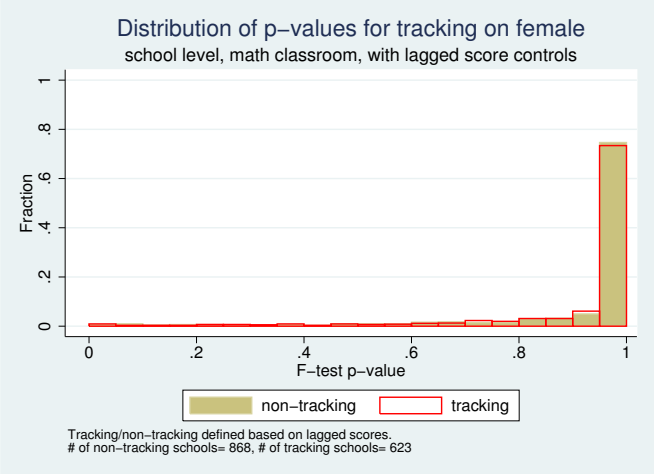
Panel B: Tracking large vs. small schools



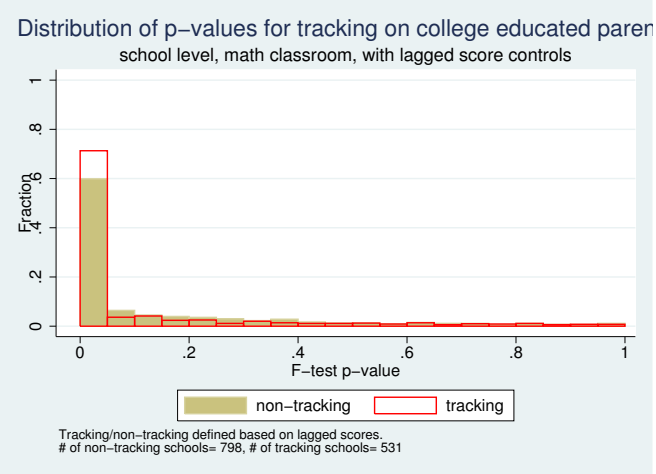
Note: Distributions of p-values are displayed from F-test on the joint significance of classroom effects within school-grade-year cells, in simulated data. test. Number of student, classroom, school and school-grade-year observations borrowed from the real data. Student-level regressions for each school, dependent variables: lagged synthetic test scores. Null hypothesis: no tracking (current classroom assignment does not predict prior achievement). Each school is represented by a p-value. Under the null hypothesis, the histogram looks like a uniform distribution (see baseline in Panel A). Schools in the first bin are classified as tracking students (see "w/ classroom component" in Panel A). Large (small) schools are defined to have above (below) median number of students from the base sample across all years. The closer the first bins are to 1 for "w/ classroom component" in Panel A and in Panel B, the more powerful the test proves to be.

Figure 5: Histogram of p-values for no tracking by other observables, controlling for lagged scores

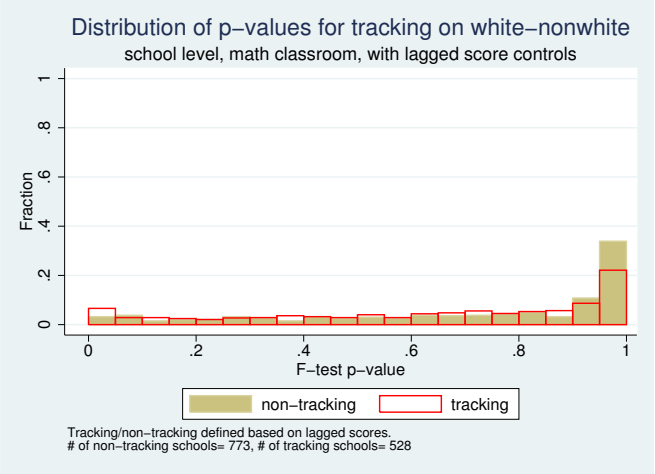
Panel A:
By Female/Male



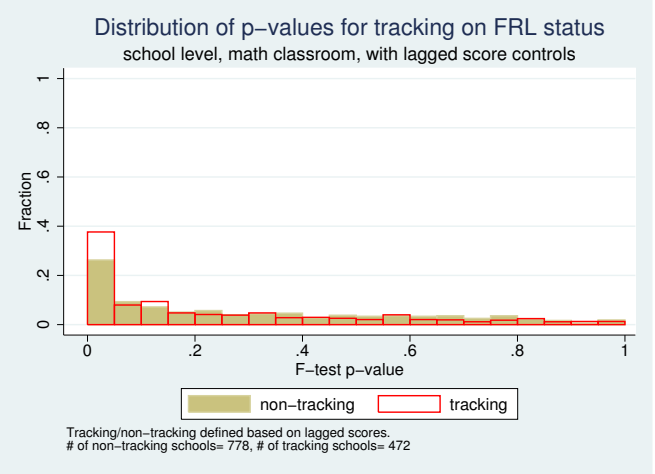
Panel C:
By College/Lower Educated Parents



Panel B:
By White/Nonwhite



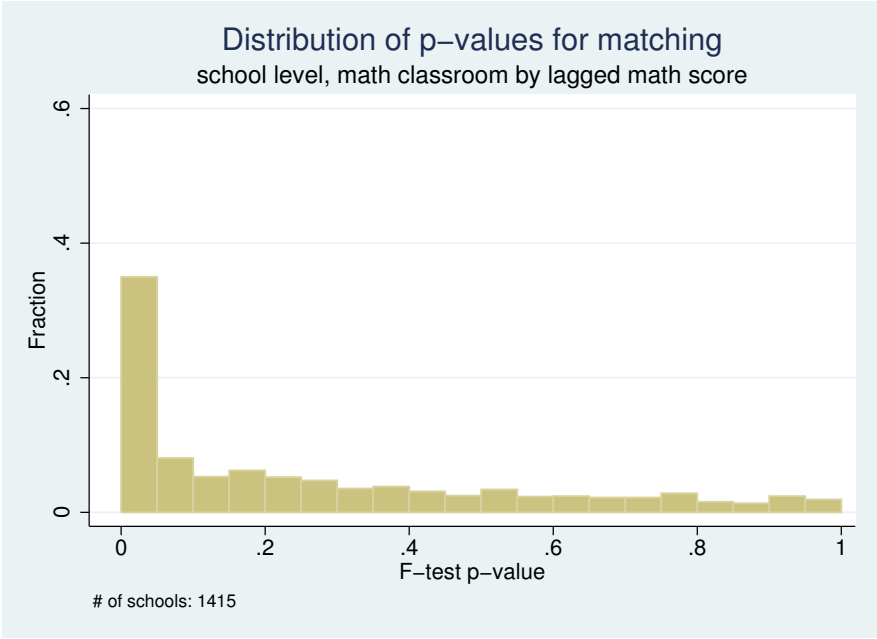
Panel D:
By FRL Eligible/Non-eligible



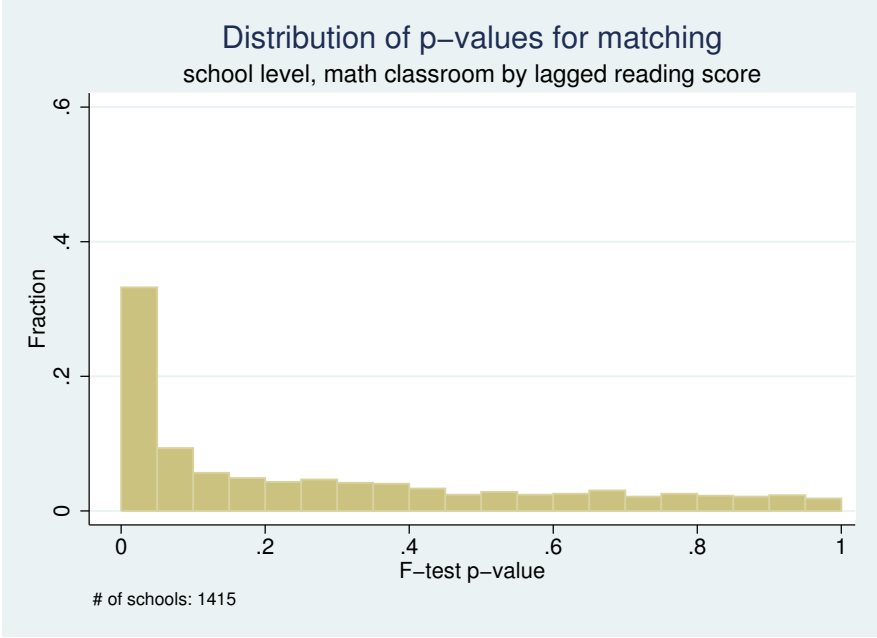
Note: Distribution of p-values are displayed from F-test on the joint significance of classroom effects within school-grade-year cells, controlling for lagged math and reading scores. Student-level regressions for each school, dependent variables: female indicator/white indicator/college educated parents indicator/FRL eligible indicator. Estimation sample for white/nonwhite (parental education/FRL) includes observations from school-grade-year cells where there is at least 15% white (college educated parent/FRL eligible) and 15% nonwhite (less than college educated parent/FRL non-eligible) students. Null hypothesis: no tracking (current classroom assignment does not predict predetermined observable characteristics). Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as tracking students. Schools in the last bin are the ones who intentionally balance their classes based on prior achievement.

Figure 6: Histogram of p-values for no matching

Panel A: By Lagged Math Scores



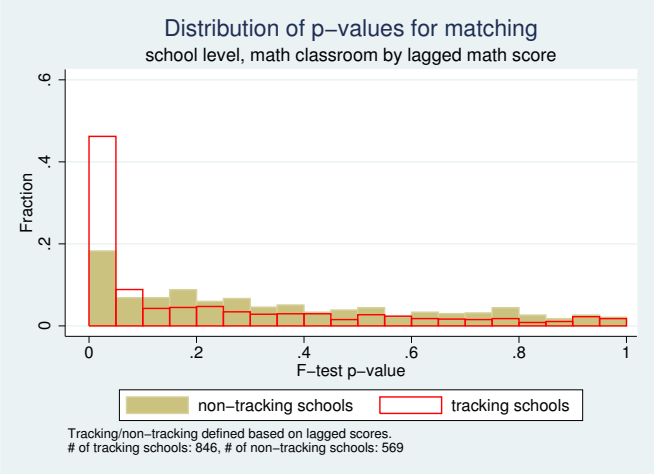
Panel B: By Lagged Reading Scores



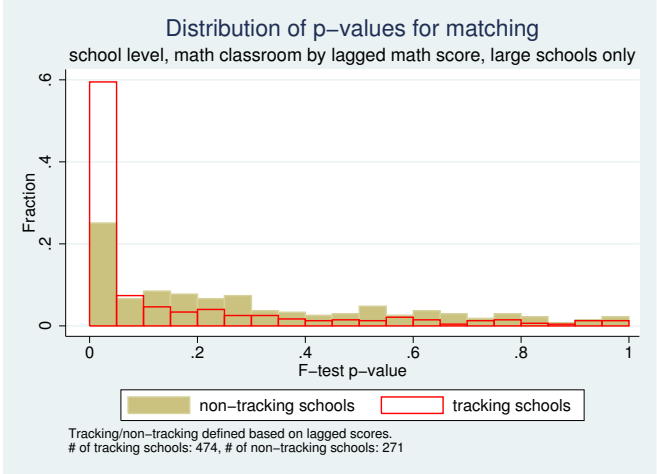
Note: Distributions of p-values are displayed from F-test on the joint significance of teacher effects within school-grade-year cells. Classroom-level regressions for each school, dependent variables: percent female/white/college educated parents/FRL eligible. Null hypothesis: no matching (current teacher assignment does not predict average prior achievement in the classroom). Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as persistently matching the same teachers to classroom of similar composition of students. Schools in the last bin would be the ones who intentionally balance their their teacher assignments across the years.

Figure 7: Histogram of p-values for no matching

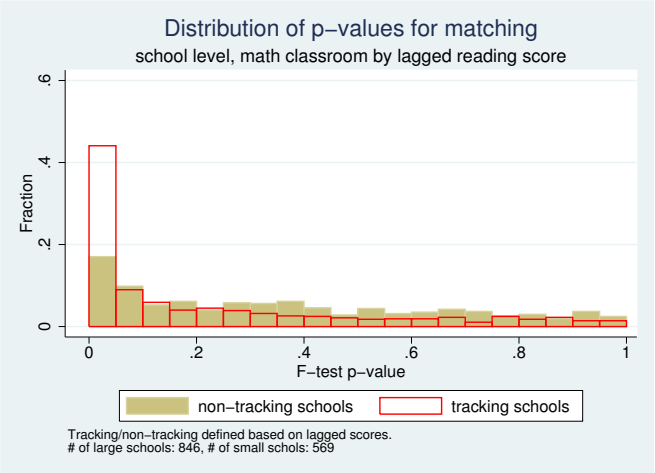
Panel A:
All Schools, By Lagged Math Scores



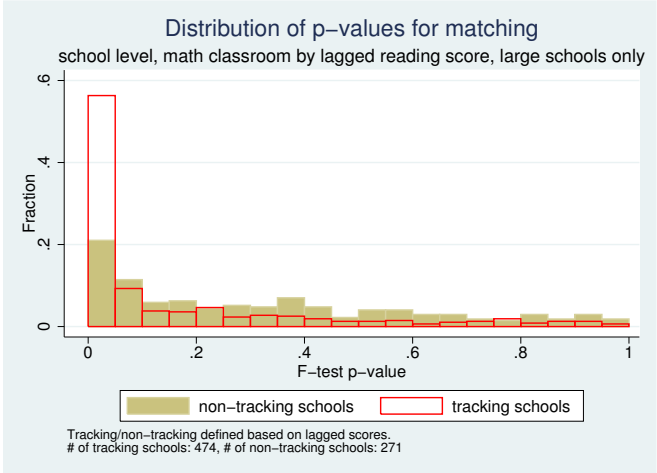
Panel C:
Large Schools, By Lagged Math Scores



Panel B:
All Schools, By Lagged Reading Scores



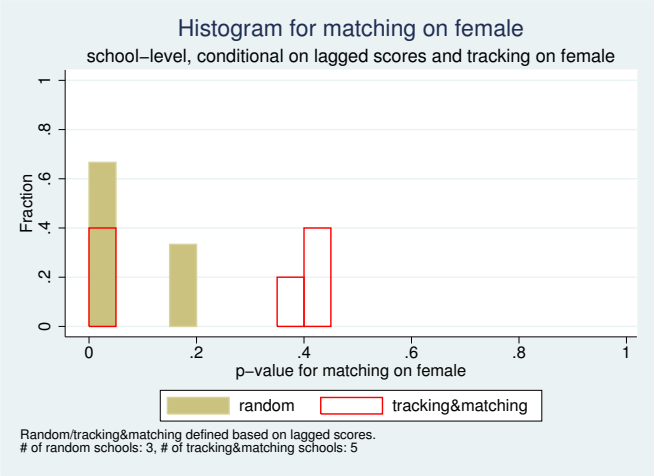
Panel D:
Large Schools, By Lagged Reading Scores



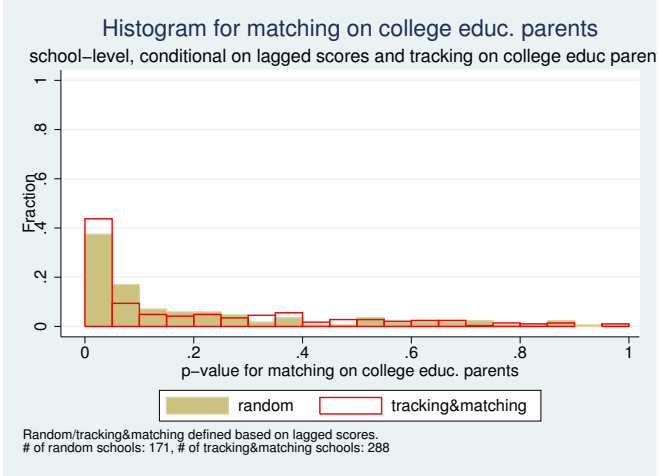
Note: Distributions of p-values are displayed from F-test on the joint significance of teacher effects within school-grade-year cells. Classroom-level regressions for each school. Null hypothesis: no matching (current teacher assignment does not predict average prior achievement in the classroom). Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as persistently matching the same teachers to classroom of similar composition of students. Schools in the last bin would be the ones who intentionally balance their their teacher assignments across the years. Large (small) schools are defined to have above (below) median number of students from the base sample across all years.

Figure 8: Histogram of p-values for no matching by other observables, controlling for lagged scores (conditional on tracking on the given variable)

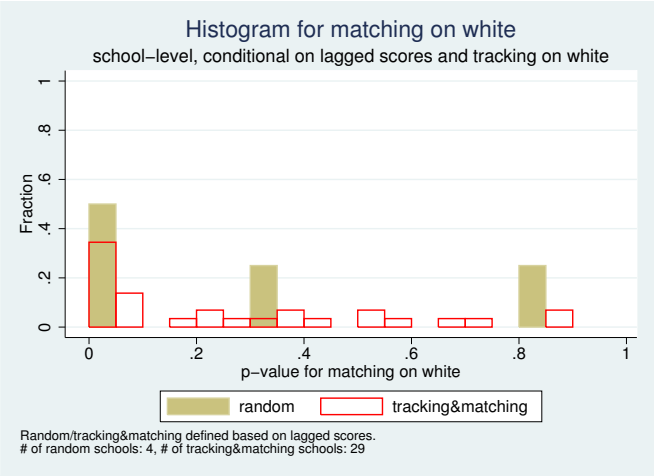
Panel A:
By Fraction Female



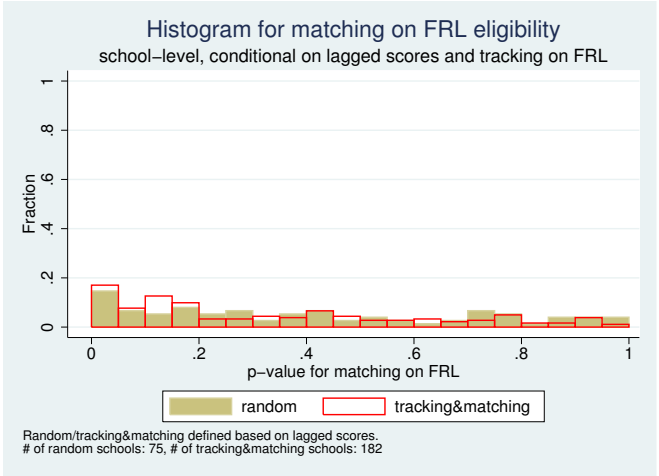
Panel C:
By Fraction with College Educated Parents



Panel B:
By Fraction White



Panel D:
By Fraction FRL Eligible



Note: Distributions of p-values are displayed from F-test on the joint significance of teacher effects within school-grade-year cells, controlling for average lagged math and reading scores. Classroom-level regressions for each school, dependent variables: percent female/white/college educated parents/FRL eligible. Estimation sample for white/nonwhite (parental education/FRL) includes observations from school-grade-year cells where there is at least 15% white (college educated parent/FRL eligible) and 15% nonwhite (less than college educated parent/FRL non-eligible) students. Null hypothesis: no matching (current teacher assignment does not predict average predetermined characteristics in the classroom). Only schools that engage in tracking based on the respective variable are shown. Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as persistently matching the same teachers to classroom of similar composition of students. Schools in the last bin would be the ones who intentionally balance their their teacher assignments across the years. Note that in Panels A-B, due to very few schools engaging in tracking based on gender or race, sample sizes are very small. There is no message to take away from these graphs, they are shown for the sake of completeness.

Appendix

A.1 Omitted Variable Bias in Teacher Effect Estimates

In this Appendix section, I formally derive the omitted variable bias when estimating the uncontrolled and partially controlled the value-added models. Let us suppress the subscripts and take the true model of test scores to be

$$y = X\beta + Z\rho + \nu, \tag{12}$$

where y is student achievement and X are student characteristics that are observed to the econometrician. On the other hand, Z are student characteristics that are unobserved for the researcher but are observed for principal. $\nu = T\mu + \theta + \varepsilon$, where T is a design matrix for teacher assignments, and θ and ε are uncorrelated with T, X and Z. Because the principal is informed about both X and Z, he/she can assign students to teachers on both. Assume that teacher assignments take the form of

$$T = X\pi_1 + Z\pi_2 + \eta, \tag{13}$$

where η is a random component. We can characterize the extent of sorting based on observables by $\pi_1 = \left(X'P_Z X\right)^{-1} X'P_Z T$, while the extent of sorting based on unobservables by $\pi_2 = \left(Z'P_X Z\right)^{-1} Z'P_X T$.[47] That is, $\pi_1 = 0$ means random sorting based on observables X, while $\pi_2 = 0$ means random sorting based on unobservables Z.

As parameters β, ρ, π_1 and π_2 are not of direct interest here, without loss of generality, we can take $\tilde{Z} \equiv P_X Z$, and rewrite eqs. (12) and (13) as

$$y = T\mu + X\left(\beta + \kappa\rho\right) + \tilde{Z}\rho + \theta + \varepsilon,$$

[47] $P_A = I - A\left(A'A\right)^{-1}A'$ denotes the orthogonal projection matrix onto A.

where $\kappa \equiv \left(X'X\right)^{-1} X'Z$; and

$$T = X\left(\pi_1 + \kappa\pi_2\right) + \tilde{Z}\pi_2 + \eta.$$

Then X is orthogonal to $\tilde{Z}$, and so $\pi_1 + \kappa\pi_2 = \left(X'X\right)^{-1} X'T$ tells us the extent of sorting X *alone*, while $\pi_2 = \left(\tilde{Z}'\tilde{Z}\right)^{-1} \tilde{Z}'T$ tells us the extent of sorting on Z after controlling for X. Similarly, $\beta + \kappa\rho$ tells us how predictive is X for test scores, while ρ says how predictive Z is *beyond* X.

Let us examine the uncontrolled model first:

$$y = T\mu + \xi, \tag{14}$$

where $\xi = X\beta + Z\rho + \theta + \varepsilon$. Then by the standard omitted variables formula,

$$\begin{aligned} OVB^{\text{uncontrolled}} &= \left(T'T\right)^{-1} T'\left[X\left(\beta + \kappa\rho\right) + \tilde{Z}\rho\right] = \left(T'T\right)^{-1}\left[T'X\left(\beta + \kappa\rho\right) + T'\tilde{Z}\rho\right] \quad (15) \\ &= \left(T'T\right)^{-1}\left[\left(\pi_1 + \kappa\pi_2\right)' X'X\left(\beta + \kappa\rho\right) + \pi_2'\tilde{Z}'\tilde{Z}\rho\right]. \quad (16) \end{aligned}$$

Note that $T'X\left(\beta + \kappa\rho\right)$ is nothing else but the index of observables, averaged over *all* students of a teacher throughout the years. Similarly, $T'\tilde{Z}\rho$ is the index of unobservables (not predicted by X), averaged again over *all* students of a teacher throughout the years. The rearranged form, in the second line of eq. (15), shows more explicitly, that $OVB^{\text{uncontrolled}}$ depends on the extents of sorting based on X and $\tilde{Z}$ (π_1 and π_2), and how well each of these predict y (β and ρ). In schools that I classify as randomly assigning, $\pi_1 + \kappa\pi_2 = 0$, therefore in these schools, omitted variable bias in the uncontrolled regression is solely coming from sorting on unobservables not predicted by X,

$$OVB_R^{\text{uncontrolled}} = \left(T'T\right)^{-1} \pi_2'\tilde{Z}'\tilde{Z}\rho,$$

where the R subscript stands for random schools.

Let us turn to the partially controlled specification now:

$$y = T\mu + X(\beta + \kappa\rho) + \tilde{\theta} + \tilde{\varepsilon}, \tag{17}$$

where $\tilde{\theta} = \overline{\tilde{Z}}\rho + \theta$ and $\tilde{\varepsilon} = \left(\tilde{Z} - \overline{\tilde{Z}}\right)\rho + \varepsilon$ and averaging takes place at the teacher-year level. The omitted variable bias in μ and $\beta + \kappa\rho$ when estimating (17) can be written as[48]

$$OVB^{\text{partially controlled}} = \begin{pmatrix} T'T & T'X \\ X'T & X'X \end{pmatrix}^{-1} \begin{pmatrix} T'\tilde{Z} \\ X'\tilde{Z} \end{pmatrix} \rho,$$

using again the standard omitted variable formula. As $X'\tilde{Z} = 0$ by construction, and by exploiting the fact that $\begin{pmatrix} T'T & T'X \\ X'T & X'X \end{pmatrix}$ is block-diagonal, we can focus on the elements that determine the bias in μ only:[49]

$$OVB^{\text{partially controlled}} = \begin{pmatrix} \left(T'P_XT\right)^{-1} & -\left(T'P_XT\right)^{-1}T'X\left(X'X\right)^{-1} \\ \cdots & \cdots \end{pmatrix} \begin{pmatrix} T'\tilde{Z} \\ 0 \end{pmatrix} \rho.$$

Therefore, the bias in the partially controlled model only comes from sorting on unobservables *beyond* X in both types of schools:

$$OVB^{\text{partially controlled}} \text{ in } \mu = \left(T'P_XT\right)^{-1}T'\tilde{Z}\rho = \left(T'P_XT\right)^{-1}\pi_2'\tilde{Z}'\tilde{Z}\rho.$$

Notice that in schools that I classify as randomly assigning, $T = P_XT$,[50] so in these schools

$$OVB^{\text{partially controlled}} = OVB^{\text{uncontrolled}} = \left(T'T\right)^{-1}\pi_2'\tilde{Z}'\tilde{Z}\rho.$$

[48] The matrix inside the inverse is non-singular as long as sorting students to teachers based on X is not perfect.

[49] Bias in $\beta + \kappa\rho$ is not of direct interest here.

[50] Since $X'T = \pi_1 + \kappa\pi_2 = 0$.

A.2 Tables

Table A.1: Simulation of Tracking with a Mix of Tracking and Non-Tracking schools: Percent Correctly Predicted and Misclassification

# schools=1,415	don't reject	reject	Total
non-tracking	0.941	0.056	1.000 (0.408)
tracking	0.078	0.922	1.000 (0.592)
Total	0.431	0.569	1.000

Note: Share of correctly and misclassified schools in the simulation of the tracking test with a mix of tracking (60%) and non-tracking (40%) schools. The simulation uses the number of student, classroom, school and school-grade-year observations from the real data. The more powerful the test is, the higher the share of correctly specified schools are (the sum of numbers in the diagonal). See more details in Section 4.1.1.

Table A.2: Correlation of Tracking by Different Observables

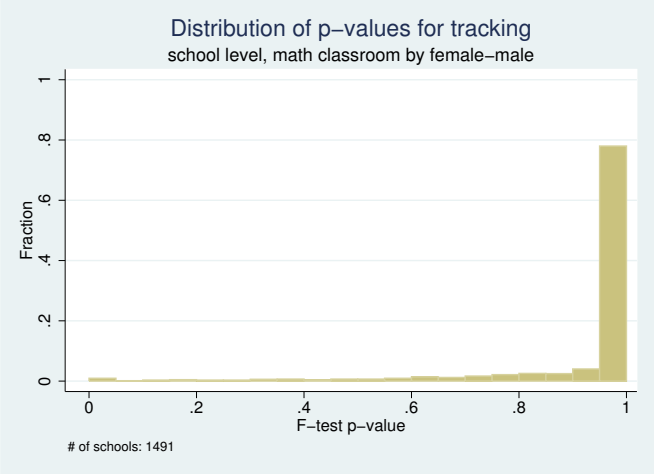
# schools=1082	lagged math	lagged reading	female	white-nonwhite	coll. educ. parents	FRL elig.
lagged math	1.000					
lagged reading	0.763	1.000				
female	0.035	0.039	1.000			
white-nonwhite	0.200	0.224	0.092	1.000		
coll. educ. parents	0.227	0.237	-0.006	0.078	1.000	
FRL eli.	0.336	0.307	0.032	0.244	0.215	1.000

Note: Correlation between tracking indicators by different observable characteristics. Higher numbers mean schools that track based on one variables, also tend to track based on the other. Correlation coefficients are weighted by school size. Number of student observations: 2,396,524. Tracking is defined as having a p-value$\leq$0.05 in the tracking F-test. In this case, the tracking F-test was performed independently for each variable, without controlling for anything else other than school-grade-year cells.

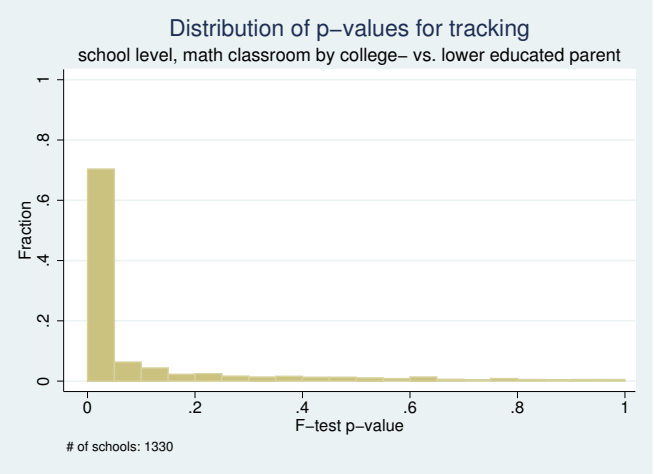
A.3 Figures

Figure A.1: Histogram of p-values for no tracking by other observables

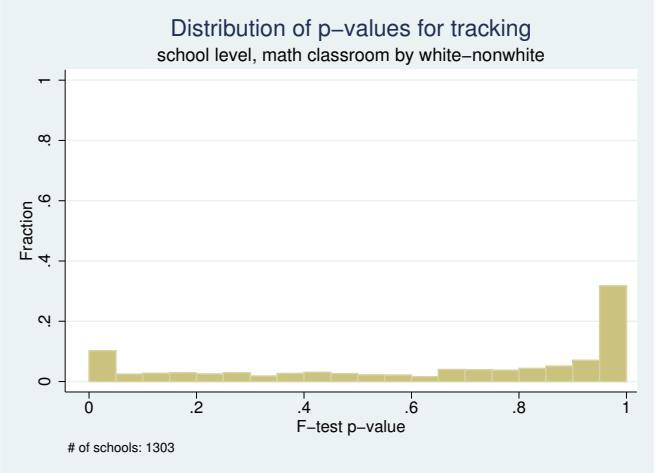
Panel A:
By Female/Male



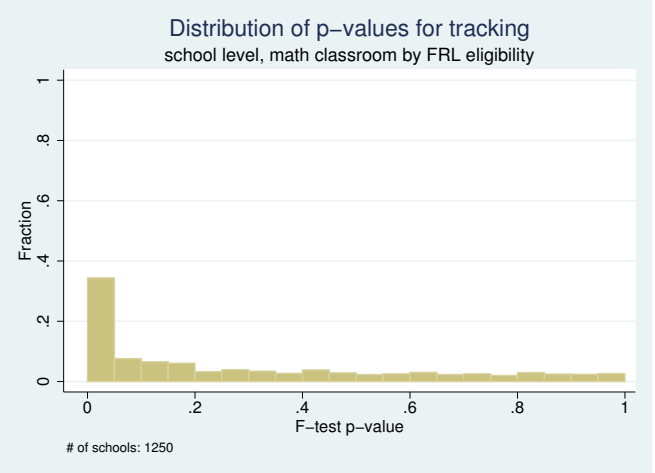
Panel C:
By College/Lower Educated Parents



Panel B:
By White/Nonwhite



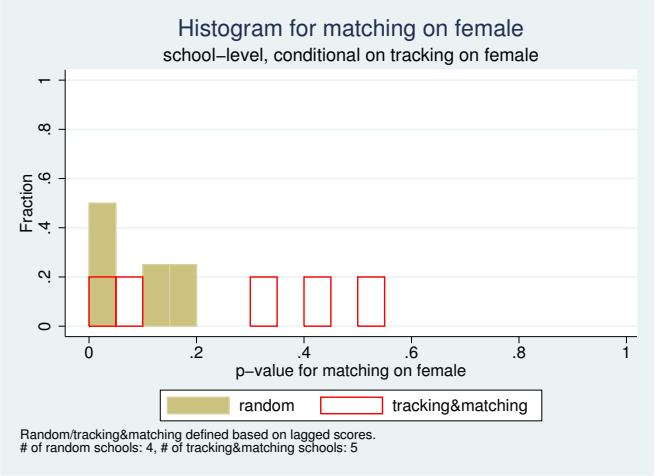
Panel D:
By FRL Eligible/Non-eligible



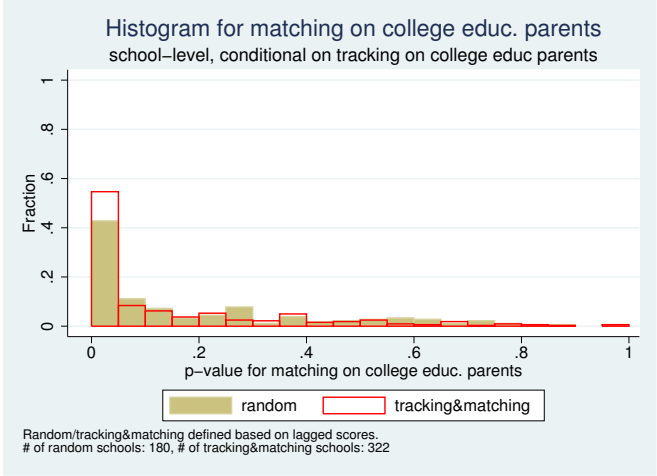
Note: Distributions of p-values are displayed from F-test on the joint significance of classroom effects within school-grade-year cells, not controlling for prior achievement. Student-level regressions for each school, dependent variables: female indicator/white indicator/college educated parents indicator/FRL eligible indicator. Estimation sample for white/nonwhite (parental education/FRL) includes observations from school-grade-year cells where there is at least 15% white (college educated parent/FRL eligible) and 15% nonwhite (less than college educated parent/FRL non-eligible) students. Null hypothesis: no tracking (current classroom assignment does not predetermined observable characteristics). Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as tracking students. Schools in the last bin are the ones who intentionally balance their classes based on the respective variable.

Figure A.2: Histogram of p-values for no matching by other observables (conditional on tracking on the given variable)

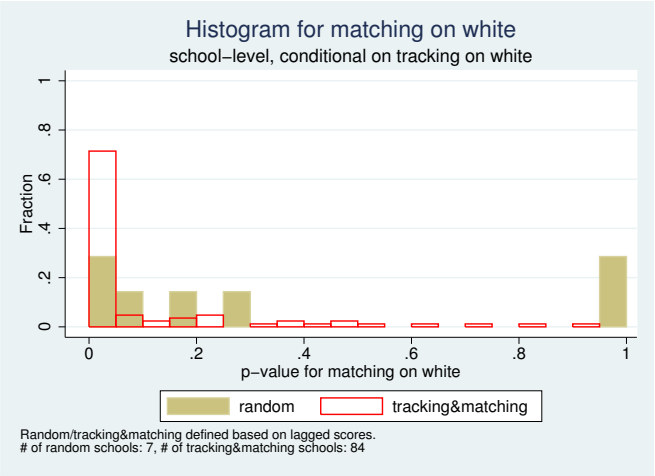
Panel A:
By Fraction Female



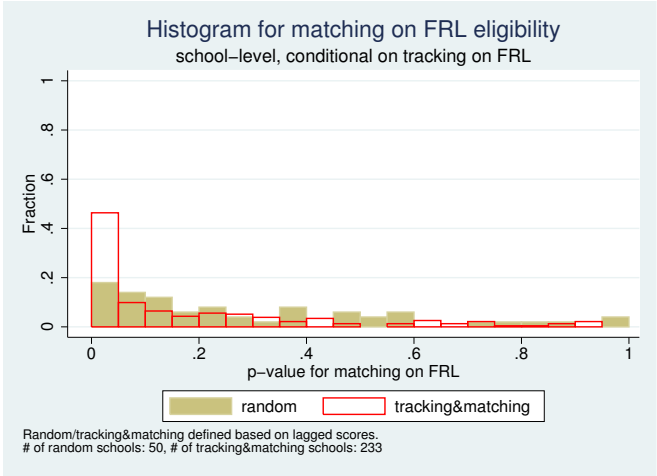
Panel C:
By Fraction with College Educated Parents



Panel B:
By Fraction White



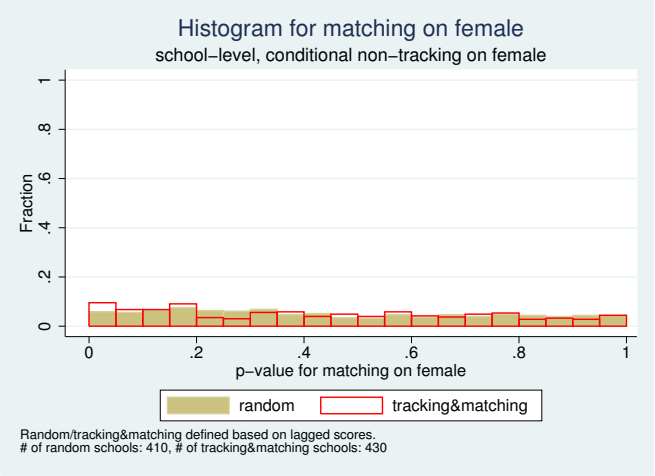
Panel D:
By Fraction FRL Eligible



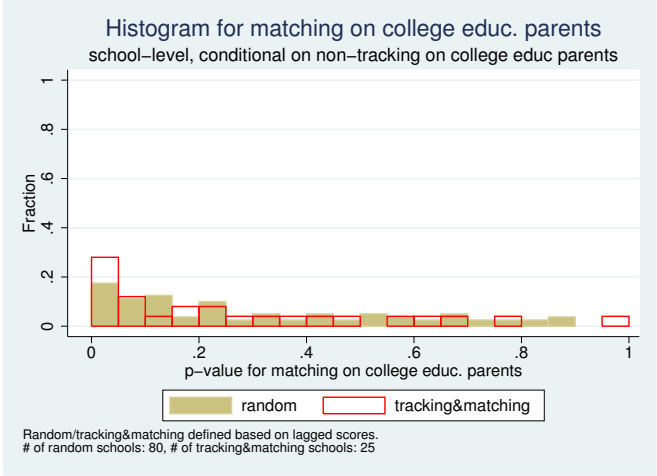
Note: Distributions of p-values are displayed from F-test on the joint significance of teacher effects within school-grade-year cells, not controlling for average prior achievement. Classroom-level regressions for each school, dependent variables: percent female/white/college educated parents/FRL eligible. Estimation sample for white/nonwhite (parental education/FRL) includes observations from school-grade-year cells where there is at least 15% white (college educated parent/FRL eligible) and 15% nonwhite (less than college educated parent/FRL non-eligible) students. Null hypothesis: no matching (current teacher assignment does not predict average predetermined characteristics in the classroom). Only schools that engage in tracking based on the respective variable are shown. Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as persistently matching the same teachers to classroom of similar composition of students. Schools in the last bin would be the ones who intentionally balance their their teacher assignments across the years. Note that in Panels A-B, due to very few schools engaging in tracking based on gender or race, sample sizes are very small. There is no message to take away from these graphs, they are shown for the sake of completeness.

Figure A.3: Histogram of p-values for no matching by other observables (conditional on non-tracking on the given variable)

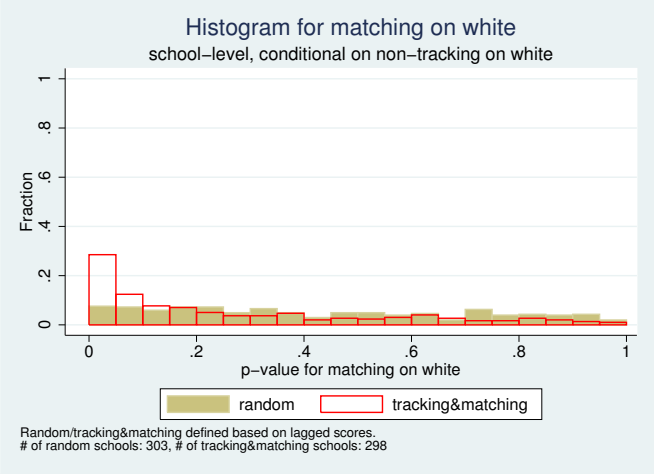
Panel A:
By Fraction Female



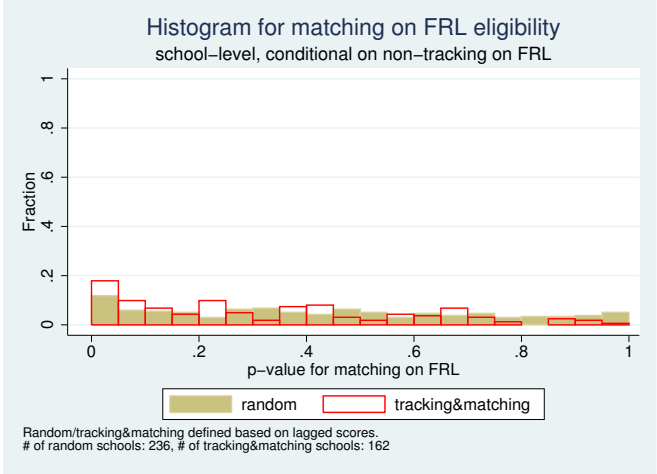
Panel C:
By Fraction with College Educated Parents



Panel B:
By Fraction White



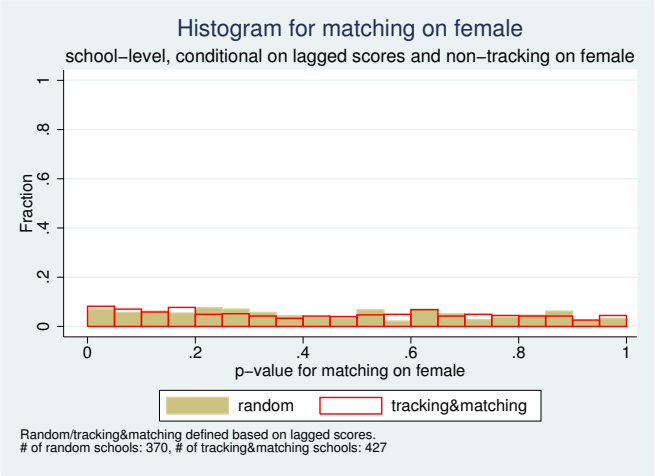
Panel D:
By Fraction FRL Eligible



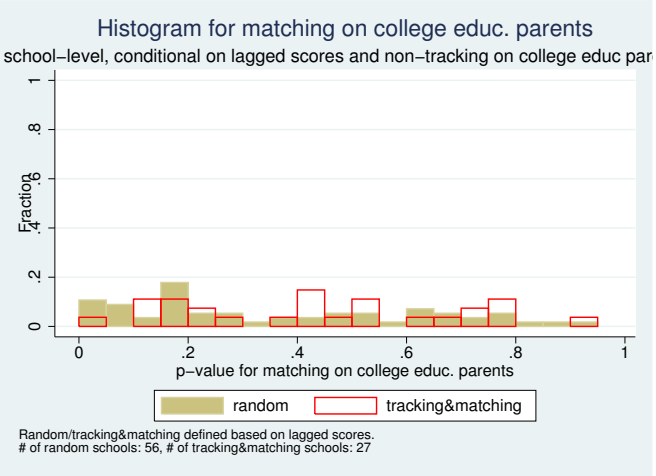
Note: Distributions of p-values are displayed from F-test on the joint significance of teacher effects within school-grade-year cells, not controlling for average prior achievement. Classroom-level regressions for each school, dependent variables: percent female/white/college educated parents/FRL eligible. Estimation sample for white/nonwhite (parental education/FRL) includes observations from school-grade-year cells where there is at least 15% white (college educated parent/FRL eligible) and 15% nonwhite (less than college educated parent/FRL non-eligible) students. Null hypothesis: no matching (current teacher assignment does not predict average predetermined characteristics in the classroom). Only schools that do not engage in tracking based on the respective variable are shown. Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as persistently matching the same teachers to classroom of similar composition of students. Schools in the last bin would be the ones who intentionally balance their their teacher assignments across the years.

Figure A.4: Histogram of p-values for no matching by other observables, controlling for lagged scores (conditional on non-tracking on the given variable)

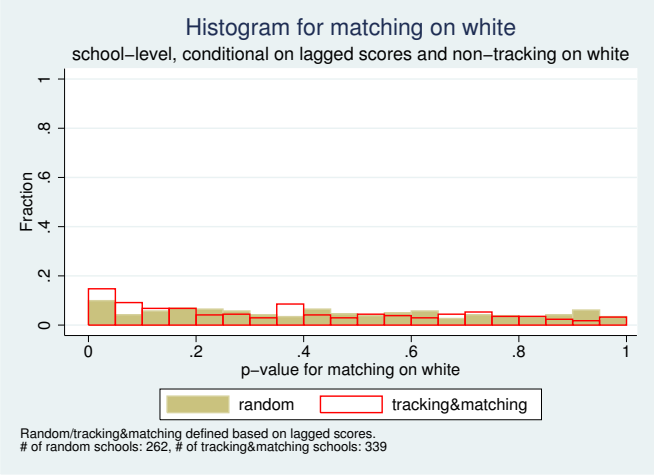
Panel A:
By Fraction Female



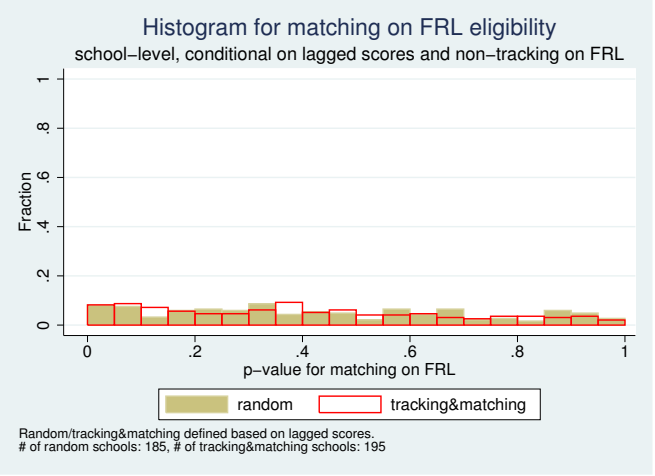
Panel C:
By Fraction with College Educated Parents



Panel B:
By Fraction White



Panel D:
By Fraction FRL Eligible



Note: Distributions of p-values are displayed from F-test on the joint significance of teacher effects within school-grade-year cells, controlling for average lagged math and reading scores. Classroom-level regressions for each school, dependent variables: percent female/white/college educated parents/FRL eligible. Estimation sample for white/nonwhite (parental education/FRL) includes observations from school-grade-year cells where there is at least 15% white (college educated parent/FRL eligible) and 15% nonwhite (less than college educated parent/FRL non-eligible) students. Null hypothesis: no matching (current teacher assignment does not predict average predetermined characteristics in the classroom). Only schools that do not engage in tracking based on the respective variable are shown. Each school is represented by a p-value. Under the null hypothesis, the histogram would look like a uniform distribution. Schools in the first bin are classified as persistently matching the same teachers to classroom of similar composition of students. Schools in the last bin would be the ones who intentionally balance their their teacher assignments across the years.