Economics of Education Review 40 (2014) 76–92

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Teacher heterogeneity, value-added and education policy☆

Scott Condie [a,*], Lars Lefgren [b], David Sims [c]

[a] 136 Faculty Office Building, Provo, UT 84602, United States
[b] 182 Faculty Office Building, Provo, UT 84602, United States
[c] 164 Faculty Office Building, Provo, UT 84602, United States

ARTICLE INFO

Article history:
Received 3 May 2013
Received in revised form 9 September 2013
Accepted 21 November 2013
Available online 8 January 2014

JEL classification:
I2
J4

Keywords:
Value-added
Teacher heterogeneity
Education policy

ABSTRACT

This study examines the theoretical and practical implications of ranking teachers with a one-dimensional value-added metric when teacher effectiveness varies across subjects or student types. We create a theoretical framework which suggests specific tests of the standard teacher input homogeneity assumption. Using North Carolina data we show that value-added fails to empirically meet these tests and document that this leads to a large number of teacher misrankings. Thus, critics of potential value-added teacher personnel policies are correct that such policies will terminate many of the wrong teachers. However, we derive the conditions under which such policies will improve student test scores and find that they will almost certainly be met. We then demonstrate that value-added information can also be used to improve student test scores by matching teachers to students or subjects according to their comparative advantage. These matching gains likely exceed those of a feasible, value-added based firing policy.

1. Introduction

In the summer of 2010, the Los Angeles Times made some news of its own by publishing an online database of value-added test scores, linked to teacher names, for thousands of teachers in the Los Angeles Unified school district. Though this generated millions of page hits for the paper's website it also produced substantial controversy about the proper use of value-added test scores in making education decisions.[1] While a few school districts, most notably the New York City schools, have explored making such teacher scores publicly available, a broader trend in many states has been the increasing use of non-publicized value-added scores in making teacher retention, promotion, and compensation decisions. Indeed, in the fall of 2012, the teachers union in the Chicago Public schools went on strike, partly to reverse the increased use of value-added test scores in teacher evaluation.

The controversy surrounding the use of value-added metrics for evaluating teacher performance largely stems from different weights given by observers to two widely known facts. First, there is substantial, measurable heterogeneity in the ability of teachers to raise student achievement. Second, value-added models depend on several assumptions, some of which have been shown to be empirically invalid in some situations.

One group of observers posits the primacy of the first fact, and emphasizes the promise of value-added to finally create an objective metric for ranking teachers and providing them with incentives. The other group emphasizes the costs of mistaken rankings and generally opposes the whole enterprise. This divergence highlights two important shortcomings of the current literature. First, while economists have produced a number of excellent statistical studies that measure the potential benefits of value-added and the distortions introduced by violating the assumptions of random teacher–student assignments,

☆ We thank Bentley MacLeod and Brian Jacob and seminar participants at Brigham Young University and the University of Michigan for helpful feedback.

* Corresponding author. Tel.: +1 8014225306.
E-mail addresses: ssc@byu.edu (S. Condie), l-lefgren@byu.edu (L. Lefgren), davesims@byu.edu (D. Sims).

[1] The teacher ratings and descriptions can be found at http://www.latimes.com/news/local/teachers-investigation/.

0272-7757/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.econedurev.2013.11.009

they have not yet integrated these with a formal model that shows how policymakers could systematically weigh costs against benefits. Second, the natural focus on statistical parameter identification has led value-added researchers to focus on between teacher heterogeneity and consequently on policies that use value-added for comparing effectiveness between teachers and rewarding success along that dimension.

Conceptually, it is not hard to imagine that one teacher may be more effective than another with one type of students, even as the ranking is reversed with a second type of students. Alternatively, one teacher may be particularly effective at teaching mathematics while another excels at teaching children to read. Such differences across teachers have been empirically documented and likely seem self-evident to parents of school-aged children. However, these concerns are generally assumed away or minimized when considering statistical methods to rank teachers. Hence, potential policy levers such as the non-random assignment of students to teachers actually come to be regarded as a nuisance because they threaten the statistical identification of the teacher quality parameter.

In this paper, we explore the promise and pitfalls of using value-added test scores to improve student learning when the traditional assumption of teacher homogeneity across groups is called into question. To do this, we outline a model of student learning that allows for teacher effectiveness to vary across subjects or student types. We next develop and, using data from 4th and 5th grade students in North Carolina, execute tests of the standard value-added assumptions about the heterogeniety of teacher ability across subjects and types of students. We find that the heterogeneity assumptions of the standard value-added model are rejected. At plausible parameter values, value-added provides a potentially misleading pairwise comparison of teachers between 15 and 25 percent of the time. These misrankings generally arise because value-added indicates one teacher to be superior to another when in fact she is better with only one subject or with one type of students. Furthermore, in the neighborhood of a policy-dictated, value-added cutoff, these measures misrank teachers relative to their social value even in the limit in a majority of cases. Hence, a teacher who is fired may generate more social value than one who is retained. Thus, part of the critique of using value-added measures for personnel policies is undoubtedly correct, such policies will unfairly fire a large number of the wrong teachers.

However, as the primary goal of using value-added personnel policies is to improve student test scores, we also derive the conditions under which this will occur and find them to be almost certainly met. Indeed, simulations with our data confirm that value-added policies would in fact improve social welfare under almost any plausible set of parameter values. Under the parameter values we actually observe, a sample policy of firing teachers in the bottom 10 percent of the observed value-added distribution would increase single-year student achievement by about 0.016 standard deviations in reading and 0.030 standard deviations in math.

Nevertheless, the misrankings of teachers under the standard value-added assumptions suggest that there might be a better way to use value-added information to improve student test scores. Indeed, we prove that assigning teachers to subjects or student types with which they match well can increase student welfare relative to random assignment. Our simulations suggest that teacher specialization according to student ability has the potential to increase reading performance of all students by approximately 0.025 standard deviations. The benefits of assigning teachers to specialize in the subject in which they excel are even larger. This policy would raise math achievement by 0.05 standard deviations and reading achievement by 0.03 standard deviations. These matching improvements are all larger than those realized by replacing the bottom 10 percent of teachers in the value-added distribution. Thus, while it raises concerns about the use of value-added in personnel policies, the heterogeneity of teacher effectiveness across subjects and settings also provides a low-cost potential policy to improve the achievement of all student types.

The remainder of the paper proceeds as described below. Section 2 provides information and highlights the most directly relevant prior research on teacher value-added models. It also describes the data we use for our analysis. Section 3 introduces our model of teacher ability rankings, which allows for teacher ability heterogeneity across student types or subjects, and shows how standard value-added measures emerge as a special case under specific assumptions. Section 4 uses our model to illuminate the conditions under which value-added measurement captures a correct ranking of teacher effectiveness, and then tests those conditions using our data. Section 5 discusses the welfare implications of using value-added to rank teachers when these conditions are not met, and presents policy simulations of the student achievement effects of using value-added for teacher de-selection versus other potential interventions. This section also explores the optimal assignment of teachers to subjects and student types. Section 6 concludes. To aid in the exposition, the proofs of all of the results are found in the appendix.

2. Background and data

2.1. Background

The rise of value-added modeling for teacher evaluation has promised a data-driven method to assess teacher effectiveness. These models leverage extensive longitudinal arrays of matched teacher–student data to produce a numerical measure of a teacher's ability to affect student achievement. Using such models Rivkin, Hanushek, and Kain (2005), Rockoff (2004), and Aaronson, Barrow, and Sander (2007) demonstrate that there are substantial, important differences across teachers in the ability to improve student achievement. This variability of teacher effectiveness has further been confirmed in a large experimental setting by Kane and Staiger (2008). These studies suggest that a one standard increase in teacher value-added ability results in a student test score increase

of 0.10–0.18 standard deviations in reading and 0.11–0.22 standard deviations in math.

Education researchers have been careful in describing the conditions under which value-added estimates produce valid rankings of teacher (or school) effectiveness. Following the discussion of Reardon and Raudenbush (2009), value-added models rely on six fundamental assumptions. (1) Manipulability of the outcome; (2) no interference between units; (3) the test scores are measured in interval units of social interest; (4) teacher effects are invariant across student types; (5) teacher assignment is conditionally independent of unobservable factors affecting achievement; and (6) the functional form of the learning model is correct.

The literature has also provided a number of critiques suggesting that these assumptions may be empirically unfounded. Rothstein (2010) provides a prominent recent critique that the dynamic selection process of students to input streams leads to inconsistent estimation. This violation of assumption 5 above is a natural extension of an earlier critique that students are not randomly assigned to teachers (Clotfelter, Ladd, & Vigdor, 2006; Monk, 1987), accompanied by evidence that the non-randomness cannot be conditioned out using student-level controls. In another sense this critique is a broadening of the Todd and Wolpin (2003) observation that fixed student characteristics (even simple fixed effects) are unlikely to fully account for the family influence on student achievement since that influence may adjust to compensate for school level resource assignments.

In response to this critique, there has been a renewed focus on the degree to which the non-random assignment of students to teachers biases value-added estimates, with some researchers arguing that the estimates square well with those from experimental assignment groups (Kane & Staiger, 2008) and another hoping that the, "assumptions are not violated too severely," (Harris, 2009). While these responses are important attempts to understand the problem, they provide only scant, non-generalizable evidence about the practical implications of the assumption violations. More recently, Guarino, Reckase, and Wooldridge (2011), attempt to simulate the degree of bias to value-added measures in different teacher assignment scenarios. However, they still give no guidance as to what level of assumption violations would be "too serious" or how such a criteria could even be formed.

Those who feel the non-random assignment violations are not too severe suggest that value-added models provide policy makers with two potential levers for increasing student achievement. First, value-added measures can be used to improve the performance of existing teachers. Prior to this study, the primary proposals to target this lever have been based on providing incentive rewards for superior teacher performance as measured by value-added. Unfortunately, there is little evidence that teachers are able to successfully identify which behaviors to change in order to increase performance (Hanushek & Rivkin, 2004; Springer et al., 2010). Second, value-added may be used to deselect bad teachers, either by firing current teachers with low value-added, or by failing to retain new teachers who demonstrate low value-added (Gordon, Kane, & Staiger, 2006; Hanushek, 2002). In this scenario, the teacher labor force will gradually improve as the worst elements are weeded out. Of course, this presumes that value-added can correctly identify the worst performing teachers and that the pool of potential replacements will be noticeably better than those they supplant.

In this study we are primarily concerned with the violation of assumption 4, the invariance of teacher effects. This assumption has received less attention in the literature with the notable exceptions of Koedel and Betts (2007) and Lockwood and McCaffrey (2009). The former look at the interaction between value-added and demographics in the San Diego school district and find no evidence of heterogeneous teacher effects. In a broader sample, and using a different methodology, the latter study finds that teacher effects are not homogenous across student ability, but that the differential teacher effects are small, explaining less than 10 percent of the overall teacher effect. While the present study adds to the evidence of these previous papers, one of its main contributions is to specify the practical consequences of violating this assumption. To do so, the paper provides a general theoretical framework to guide the empirical investigation and derive the implications of within-teacher skill heterogeneity. Without such a theory, it is hard to specify the circumstances under which teacher personnel policies based on value-added models will improve student welfare.

Our focus also suggests a different type of policy option. In particular, we could use value-added information to improve the performance of current teachers by altering the matching between students and teachers. Thus, instead of viewing non-random assignment as the impediment to the use of value-added, it actually becomes a key mechanism through which value-added measures can improve student achievement.

2.2. Data

The empirical exercises in this paper use a data set derived from North Carolina school administrative records maintained by the North Carolina Education Research Data Center. The student-year observations for grades 4–5 are taken from the years 1998–2004, and provide score information from standardized tests in reading and mathematics as well as basic student demographic characteristics. We scale the test scores to reflect standard-deviation units relative to the state average for that grade and year. These student level records are then matched to personnel record data for classroom teachers. Our final data set reflects a slightly greater than 75% match success rate between students and teachers, which approximates the success rate obtained by previous users of this data (Clotfelter, Ladd, & Vigdor, 2007; Jacob, Lefgren, & Sims, 2010).

Table 1 reports summary statistics including the basic demographic controls for student race, ethnicity, free lunch and special education status available in the data. While the North Carolina sample is close to the national average in free lunch eligibility (45 percent compared to 42 percent nationally) it actually has smaller than average minority enrollments, comprised mainly of African-American students, and has only a small percentage of non-native English speakers. We designate students as high initial ability if they

Table 1
Summary statistics for sample of students.

Variable	Mean (standard deviation)
Normalized reading score	0.000 (1.000)
Normalized math score	0.000 (1.000)
Student fraction male	0.506 (0.500)
Student fraction free lunch	0.451 (0.498)
Student fraction white	0.618 (0.486)
Student fraction black	0.292 (0.455)
Student fraction Hispanic	0.045 (0.208)
Student fraction special ed.	0.127 (0.333)
Student fraction limited English	0.028 (0.165)
Student age	10.754 (0.721)
Parent education < high school	0.116 (0.320)
Parent education–high school	0.198 (0.398)
Parent education–college grad	0.209 (0.406)
Parent education–graduate work	0.048 (0.214)

Test scores are normalized to be mean zero with unit standard deviation by state/district, year, grade, and subject. There are n = 1,323,964 students in the sample.

are in the top third of their school-grade ability distribution based on their previous year test score, and low initial ability if they are in the bottom third of the distribution.

To test the teacher heterogeneity assumption in value-added models (assumption 4 above), we calculate value-added measures of teachers for both reading and math for the full sample of students. We also construct separate value-added measures for high-ability and low-ability students.[2] We are aware of the potential limitations of these calculations with respect to other maintained assumptions, for example about test score scaling and rates of decay, though a detailed examination of all possible violations is beyond the scope of the present study. In this paper we have chosen to focus on the assumption of homogenous teacher effects.

In constructing our empirical measures of value-added, we assume that student learning A_i evolves according to the following process:

$$A_{ijsky} = test_{ijsky} - test_{ijsky-1} = X_{isy}\Gamma + \theta_{jsk} + \eta_{jsky} + \epsilon_{ijsk} \quad (2.1)$$

where i indexes a student in the classroom of a specific teacher j, in school s, taking subject k, in year y. Note that in this empirical framework, we have indicated teacher effectiveness as θ_{jsk}. In the context of our theoretical model described below, this corresponds to $\mu_j\nu$, or the amount of teacher input multiplied by the student's sensitivity to teaching inputs, both of which we can further allow to vary by student type $t(i)$ or subject k. In this framework, an average student level residual from a classroom-year is interpreted as the estimate of teacher effectiveness $\hat{V}_{jsky} = \theta_{jsk} + \eta_{jsky} + (1/N)\sum_{i=1}^{N}\epsilon_{ijsky}$.

In practice we run year by year regressions of achievement gains on a set of student and school level covariates. We estimate separate value-added measures for each subject using the entire sample. We then run separate regressions using just high-ability or low-ability students. Our estimates of teacher effectiveness reflect both sampling variation and potential classroom or school specific shocks.[3] If we assume, however, that shocks are uncorrelated over time, we can identify moments of the true distribution of teacher effectiveness. More specifically, it is the case that $Var(V_{js}^{t}) = Cov(\hat{V}_{jsy}^{t}, \hat{V}_{jsy-1}^{t})$ and $Cov(V_{js}^{t=1}, V_{js}^{t=2}) = Cov(\hat{V}_{jsy}^{t=1}, \hat{V}_{jsy-1}^{t=2})$, where t indexes either subject or student type.

Although we calculate teacher value-added using a gains procedure (differencing sequential years' test scores), alternative models using a specification employing lagged past-year test scores produce similar results. Other specification checks, such as making the value-added measures within school by including school-by-year grade effects actually produce lower correlations for teacher value-added across both student types and subjects suggesting that our approach is relatively conservative.

3. A theoretical framework for describing teacher value-added

This section describes a general analytical framework for examining teacher quality which nests the value-added assumptions listed in Section 2. Let teachers be enumerated by $j \in \{1, \ldots, J\}$ and students by $i \in \{1, \ldots I\}$. Each student i has a type $t(i) \in \{1, \ldots, T\}$. Teachers may have differential skills in teaching math (m) and reading (r) and across different student types. The marginal impact of teacher j on the achievement of a student i in subject $k \in \{m, r\}$ is $\mu_{t(i)jk}$. The sensitivity to teacher inputs of student i with type $t(i)$ in subject k is given by $\nu_{t(i)k}$. Consider student i who is instructed by teacher j. For each subject $k \in \{m, r\}$, achievement A_{ik} is determined by the function

$$A_{ik} = \mu_{t(i)jk}\nu_{t(i)k} + \epsilon_{ik} \quad (3.1)$$

where ϵ_{ik} is a student and subject specific error term. Consistent with value-added assumption 5, we assume

[2] Splitting the sample by initial ability also allows for differential mean reversion across student types.

[3] To help understand how our estimates of teacher ability might be influenced by other sources of variation in our test score data, we have computed one- and five-year reliability ratios. For reading, the one-year ratios are about 0.28 and the five-year measures are 0.66. In math the respective ratios are 0.52 and 0.84. More importantly there is no meaningful difference either across grades or across high-ability and low-ability students within subjects in test reliability.

throughout the paper that these error terms are independent across students and that the variances of these error terms are uniformly bounded across all students and subjects.[4] The fact that achievement relies only on teacher inputs and an independent error term is also consistent with assumptions 1, 2, and 6. Assumptions 3 and 4 are true in special cases of our model, which we highlight later.

The achievement of student i, A_i, is a weighted average of her achievement in math and reading and takes the form

$$A_i = N_m A_{im} + N_r A_{ir} \tag{3.2}$$

where N_m and N_r are the weights put on math and reading in value-added measurement. We assume that $N_m > 0$, $N_r > 0$ and $N_m + N_r = 1$. The measure of teacher value-added is the average achievement of students taught by the teacher,

$$V_j = \frac{1}{I_j}\sum_{i=1}^{I_j} A_i. \tag{3.3}$$

where I_j is defined as the number of students assigned to teacher j.

For the parsimonious model considered here, (3.3) is the analog to the value-added measures used in the empirical value-added literature. This model and value-added measure could be augmented with additional covariates but the tenor of the theoretical results would remain.

To isolate the role of teacher heterogeneity across subjects and student types we consider two classes of restrictions on this general framework. In one set of models, labeled *subject heterogeneity models*, teachers are assumed to provide the same input to students of all types. We then test to see if teacher inputs differ across subjects. The second set of models, labeled *type heterogeneity models*, assume that teachers provide the same inputs across subjects. Given this restriction, we test to see if teacher inputs vary across student types. Determining the extent to which these models hold provides insight into the use of value-added measures of teacher quality in curriculum and personnel decisions.

Societal preferences over student achievement are represented by a social welfare function. While student achievement can be affected by the quality of teacher inputs, it has a random component representing unobserved student heterogeneity, time-varying unobservables or other sources of sampling variation. We define societal preferences over student achievement in the ex-ante sense, i.e. before this unexplainable variation occurs.[5] For expositional ease, we consider social preferences that are linear in expected student achievement. Therefore, these results incorporate the first-order effects that will occur in more general models of social preferences when these preferences are of the weighted utilitarian form.[6] Throughout the analysis we make the following assumption on social preferences.

Assumption 1. Social welfare is a linear function of the expected achievement of each student. This implies a social welfare function of the form

$$W = E\sum_{i=1}^{I}\sum_{k\in\{m,r\}} M_{t(i)k}A_{ik} \tag{3.4}$$

Furthermore, social preferences are increasing in the achievement of every student; that is $M_{t(i)k} > 0$ for all $t \in \mathcal{T}$.

The constants $\{M_{tk}\}_{t\in\mathcal{T},k\in\{m,r\}}$ represent the relative social value assigned to increases in expected achievement by a student of type t in subject k. These weights are important in understanding the relationship between the statistical properties of value-added measurement and the social benefits and costs of using value-added measurement in school policy decisions.

4. Testing for within-teacher heterogeneity

In this section we introduce commonly employed restrictions on the general theoretical framework and derive statistical properties that result from these restrictions. We then test the validity of these restrictions.

4.1. Subject heterogeneity models

The model presented in this section provides a framework for better understanding the social benefits of teachers that may have heterogeneous abilities across subjects. The achievement of younger students is most commonly measured across the broad subject areas of math and reading. Generally, researchers have paid little attention to the socially optimal weights for these subjects when considering teacher effectiveness. Instead, recent research typically examines a subject in isolation or imposes equal weights on each subject without explicit justification.[7] Understanding the relative strengths of a teacher in these two broad areas may lead to better personnel policy decisions.

In subject-heterogeneity models all student types are assumed to have the same sensitivity to teacher inputs, although this sensitivity may differ across subjects. Formally, we assume that $\nu_{tm} = \nu_{t'm}$ and $\nu_{tr} = \nu_{t'r}$ for all t, t', although ν_{tm} need not equal ν_{tr}. We label this common student sensitivity for subject k, ν_k. Furthermore, we assume that $\mu_{jkt} = \mu_{jkt'}$ for all subjects k and type pairs t, t'.

[4] We need not assume that the error terms are independent across subjects for a given student.

[5] As is well understood from decision theory, the existence of a functional mapping student achievement into the real line implies that societal preferences over student achievement are complete and transitive. The discussion on the existence of social welfare functions and their usefulness in studying policy is extensive. See Suzumura (2002) for an introduction to the topic.

[6] If social preferences are not of this form then more care must be taken in interpreting the results here, although we believe that these results remain useful.

[7] Ballou, Sanders, and Wright (2004) as well as Lefgren and Sims (2012) optimally weight teacher performance across subjects to obtain more efficient estimates of subject and overall value-added. However, they do not consider optimal weighting from a social welfare point of view.

We then label the amount of teacher j's input provided in subject k as μ_{jk}. Student achievement A_i can now be re-written

$$A_i = N_m A_{im} + N_r A_{ir} = N_m(\mu_{jm}\nu_m + \epsilon_{im}) + N_r(\mu_{jr}\nu_r + \epsilon_{ir}). \quad (4.1)$$

An *environment* is a set of teachers in a particular school. The models considered hereafter come from restrictions on the set of possible environments. These restrictions involve assumptions about teacher effectiveness across subject areas.

Definition 1. An environment satisfies:

(1) the *strong locational model* (SLM) if for each teacher j, $\mu_{jm} = \mu_{jr}$ and $\nu_m = \nu_r$,
(2) the *weak locational model* (WLM) if for each teacher j, $\mu_{jm} = \mu_{jr}$.
(3) the *non-locational model* (NLM) if there exists a teacher j such that $\mu_{jm} \neq \mu_{jr}$.

In other words, the set of teachers satisfies the weak-locational model if each teacher provides identical inputs across subject areas, though teacher j need not provide the same amount of inputs in both subjects as teacher j'. If the WLM holds, the ranking of teacher effectiveness across subjects is identical. The environment satisfies the strong-locational model if it satisfies the WLM and student sensitivities across the subject areas are identical. When the SLM holds, not only is the ranking of teachers across two subjects identical, but having a better teacher induces the same achievement improvement in both subjects. An environment that fails to satisfy the weak-locational model is referred to as non-locational. In such an environment, a teacher who is more effective than a colleague in one subject need not be more effective in another.

An environment that satisfies the WLM embues calculations of value-added with desirable statistical properties in subject heterogeneity models. In particular, given sufficient data the value-added statistic given in (3.3) will uncover the correct natural ordering over teacher quality that holds in the WLM. Therefore, under the WLM any incorrect ranking of teachers that is based on value-added will be due solely to sampling error, as opposed to fundamental flaws in the measure of teacher quality. For the formal statement of this result recall that student achievement A_i is a convex combination of achievement in math and reading and that N_m is the weight placed on math in that convex combination.

Lemma 1. *Suppose that in a subject heterogeneity model the environment satisfies the WLM. Then*

$$P\left\{\lim_{I_j\to\infty} V(I_j) > \lim_{I_{j'}\to\infty} V(I_{j'})\right\} = 1 \quad (4.2)$$

for any relative subject weight N_m, if and only if $\mu_j > \mu_{j'}$.

4.2. Type-heterogeneity models

In type heterogeneity models we assume an absence of achievement-inducing heterogeneity in teacher inputs across subjects. Thus, differences in student achievement are due to differences in teacher inputs across student types and differing student type sensitivities to teacher inputs.

Under these assumptions, we can rewrite student achievement (3.1) as

$$A_i = \mu_{jt(i)}\nu_{t(i)} + \epsilon_i \quad (4.3)$$

where $\epsilon_i = N_m\epsilon_{im} + N_r\epsilon_{ir}$.

As in subject heterogeneity models, an *environment* is a set of teachers and students in a particular school. Models are assumptions about teacher effectiveness across student types and student sensitivities to teacher input.

Definition 2. An environment satisfies:

(1) the *strong locational model* (SLM) if for each teacher j, $\mu_{jt} = \mu_{jt'}$ for all t, t', and $\nu_t = \nu_{t'}$ for all t, t',
(2) the *weak locational model* (WLM) if for each teacher j, $\mu_{jt} = \mu_{jt'}$ for all t, t'.
(3) the *non-locational model* (NLM) if there exists a teacher j, such that $\mu_{jt} \neq \mu_{jt'}$ for some t, t'.

As in subject heterogeneity models, an environment that satisfies either the SLM or the WLM has a natural ordering in terms of teacher quality. Under either model a teacher j is better than another teacher j' if $\mu_{jt} > \mu_{j't}$ for all t. We denote this ordering by $\gg$. As before, an environment that satisfies the SLM also satisfies the WLM, and we classify an environment that fails to satisfy the weak locational model as a *non-locational model*.

The SLM satisfies all standard value-added assumptions including assumption 3, test scores are measured in interval units of social interest, and assumption 4, teacher effects are invariant across student types. The WLM fails to satisfy assumption 4 but satisfies a weaker assumption of monotonicity that has been discussed in the literature (Reardon & Raudenbush, 2009). Under this assumption, if teacher A is better than teacher B with one student type, teacher A is better than teacher B with all student types. Under the NLM, a teacher may provide more inputs to a particular type of student than another teacher. This teacher ranking may be reversed, however, with a different type of student.

Understanding the statistical properties of the estimator V_j from Eq. (3.3) in type-heterogeneity models requires us to define the method by which students are assigned to teachers. We therefore define an assignment function to represent the fraction of teacher j's students of each type. We assume that assignment functions do not vary with class size, although we could generalize to allow for this. More formally:

Definition 3. An assignment function $Q_j : \mathcal{T} \to [0, 1]$ defines, for a given teacher j the fraction $Q_j(t)$ of j's students that are of type t.

It follows that the value-added estimator V_j is a useful measure because it provides a consistent way to estimate the natural ordering $\gg$ over teachers under the SLM.

Lemma 2. *Suppose that the environment satisfies the SLM. Then*

$$P\left\{\lim_{I_j\to\infty} V(I_j) > \lim_{I_{j'}\to\infty} V(I_{j'})\right\} = 1 \tag{4.4}$$

for all pairs of assignment functions $Q_j, Q_{j'}$ if and only if $j \gg j'$.

This result demonstrates the desirable statistical properties that manifest themselves under the SLM in type-heterogeneity models. While analogous to Lemma 1, this result requires a stronger condition on the environment. For type-heterogeneity models, given sufficient data and the assumptions of the SLM, the value-added estimator will correctly order two teachers in terms of quality. Therefore, in type-heterogeneity models under the SLM, teacher rankings are asymptotically almost surely correct.

While the SLM is sufficient for these statistical properties to hold, we next show it is not necessary. In particular, if the student–teacher assignment function is universal (identical for all teachers), then the value-added estimator still recovers the natural ranking of teachers implied by the WLM. However, if students are matched to teachers using heterogeneous match functions then the value-added estimator need not recover this ordering even in large samples.

Lemma 3.

(1) *If the WLM holds and $Q_j \equiv Q_{j'}$ for all $j, j' \in \mathcal{J}$ then*

$$P\left\{\lim_{I_j\to\infty} V_j > \lim_{I_{j'}\to\infty} V_{j'}\right\} = 1 \tag{4.5}$$

for any j and j' for which $j \gg j'$.

(2) *If the WLM holds and $j \gg j'$ but $\max_t \mu_{j't} > \min_t \mu_{jt}$, then there exists Q_j and $Q_{j'}$ such that*

$$P\left\{\lim_{I_{j'}\to\infty} V_{j'} > \lim_{I_j\to\infty} V_j\right\} = 1 \tag{4.6}$$

This result implies that even large sample estimation of teacher value-added that does not account for student type-varying teacher assignments may yield a misleading ranking of teacher quality. To understand the intuition underlying this insight, consider the simple example of two teachers, *A* and *B*. The amount of teaching input provided by teacher *A* to all student types, μ_A, is 2 while the corresponding amount for teacher *B*, μ_B, is 1. Suppose further that there are two types of students, *C* and *D*. The elasticity of the type *C* students to teacher inputs, ν_C, is 4 while the elasticity of type *D* students to teacher inputs, ν_D, is 1. In this example, if teacher *A* was assigned all type *D* students while teacher *B* was assigned all type *C* students, teacher *A*'s value-added would converge to 2 while teacher *B*'s value-added would converge to 4. In this case, the value-added ranking reverses the underlying quality ranking of the teachers. If, however, both teachers have a common proportion of each type of student, the value-added ranking will correspond to the underlying ranking of teacher effectiveness.[8] If the types of students are observable a re-weighted value-added estimator that circumvents the bias imposed by a non-constant weighting function could alleviate this problem. Alternatively, researchers could include interactions between student type and teacher value-added as suggested by Reardon and Raudenbush (2009), though this is rarely done in the literature.

Thus, in type-heterogeneity models the commonly used value-added estimator is useful for recovering the latent teacher quality ordering when either the assumptions of the strong locational model hold or the weak locational model holds and is coupled with a universal assignment function. In a weak locational model with a non-universal assignment function or a non-locational model, value-added is a conceptually flawed measure of teacher effectiveness. We explore the policy importance of this conceptual limitation later in the paper.

4.3. Tests of the WLM and SLM

The SLM and WLM impose testable restrictions on the distribution of teacher value-added estimates V_j in both subject- and type-heterogeneity models. In particular, since the marginal impact of a teacher on student achievement is independent of either student type or subject under the SLM, the distribution of teacher value-added should be the same when measured across these categories. To develop this point further, consider calculating the value-added estimator V_j using only observations from a particular subset of data u. In type-heterogeneity models this would be a student type and in subject-heterogeneity models this would be one subject. Let I_{ju} be the cardinality of the set of students assigned to teacher j in subset u. Then we define

$$V_j(u) = \frac{1}{I_{ju}} \sum_{i=1}^{I_{ju}} A_i(u) \tag{4.7}$$

to be the value-added estimator calculated using only data from subset u.

We focus on testing the implications of the WLM and SLM for the second moments of the distribution of $V_j(u)$. Under both models, the correlation of teacher value-added across student types or subjects approaches one as the sample size approaches infinity. This is because teachers provide the same amount of educational inputs to all student types or subjects. Hence, the ranking of teacher effectiveness is always the same. Under the SLM, the variance of teacher value-added conditional on the subset u is constant for all student types or subjects. Because teachers provide the same amount of inputs and the responsiveness of achievement to inputs is the same, a teacher's observed value-added is also the same across types or subjects. Hence, the variance of value-added will also be identical.

[8] We thank Bentley MacLeod for suggesting this example.

Table 2
Value-added by subject—correlations and standard deviations.

	4th grade	5th grade
Math standard deviation	0.174**	0.177**
	(0.002)	(0.002)
Reading standard deviation	0.097**	0.098**
	(0.002)	(0.002)
Difference	0.077**	0.079*
	(0.002)	(0.002)
Correlation across subjects	0.733**	0.716**
	(0.012)	(0.015)

The estimation procedure for these standard deviations and correlations is described in Section 2.2. Standard errors are shown in parentheses. The significance of the correlation is tested against a null hypothesis of 1.
* Significance at the 10 percent level.
** Significance at the 5 percent level.

Implication 1. Let $\hat{\sigma}^2(V_j(u))$ be the sample variance of $V_j(u)$. Under the WLM,

$$\lim_{l_{ju},l_{ju'}\to\infty} P\{Corr(V_j(u),V_j(u'))-1\geq\epsilon\}=0. \tag{4.8}$$

Furthermore under the SLM

$$\lim_{l_{jt}\to\infty} P\left\{\hat{\sigma}^2(V_j(u))-\hat{\sigma}^2(V_j(u'))\geq\epsilon\right\}=0 \tag{4.9}$$

for all groups u and u'.

4.4. Testing the WLM and SLM

Table 2 shows the results of tests of the above implication for subject heterogeneity models using our North Carolina data. In particular, it first examines the second moments of the empirical teacher value-added distribution in reading and math for the entire sample of students. The results are further broken down by grade level. We identify these moments using the method described in our data section in which we examine the covariance of one year empirical value-added measures across adjacent years. Doing so allows us to estimate the true variances and covariances of the relevant teacher value-added distributions by purging them of the effects of test measurement errors.

We find that for both fourth and fifth grade students, the standard deviation of teacher value-added in mathematics is above 0.17 while the standard deviation in reading is close to 0.10. Both estimates are similar to those found in prior studies (Aaronson et al., 2007; Rivkin et al., 2005; Rockoff, 2004). This statistically significant difference between reading and math measures represents a material rejection of the strong locational model. The result suggests that student achievement is more responsive to teacher inputs in math than in reading. However, as long as the weighting of the two subjects is the same for each teacher, the ranking of teachers will be unaffected by the deviation from the strong locational model.

Consequently, for the purposes of ranking teachers, it is more critical to test whether teacher value-added is perfectly correlated across subjects. Examining the final row of Table 2, we see that the correlation across subjects is just above 0.7. This suggests that while teacher ability is correlated across subjects, there will exist teachers who are relatively more effective at teaching math than reading and vice versa. This also represents a material rejection of both the strong and weak locational model in favor of the non-locational model.

In Table 3, we test the implications of type heterogeneity model environments, letting student type correspond to initial ability level. We do so separately for reading and math. Examining the third row of the table, we see that the standard deviation of teacher value-added is virtually identical, whether examining students with low or high initial ability. While some of the differences are statistically significant, they are not practically significant. This suggests that both high-ability and low-ability students are equally responsive to the quantity of teacher provided learning inputs. This test is supportive of the strong locational model, at least in the context of a single subject. The final row of Table 3 shows the correlation of teacher value-added across students of high and low ability. For math, the correlations exceed 0.97 but are still significantly different from 1. While these represent a statistical rejection of the strong locational model, the small differences in magnitude from a perfect correlation suggest that the material impact of the rejection is likely to be quite small. For reading, however, the correlations between value-added with low-ability and

Table 3
Value-added by initial student ability—correlations and standard deviations.

	4th grade Math	5th grade Math	4th grade Reading	5th grade Reading
High-ability standard deviation	0.176**	0.178**	0.105**	0.102**
	(0.002)	(0.002)	(0.002)	(0.002)
Low-ability standard deviation	0.176**	0.178**	0.105**	0.107**
	(0.002)	(0.002)	(0.002)	(0.002)
Difference	−0.001	0.000*	0.000	−0.006**
	(0.000)	(0.001)	(0.001)	(0.001)
Correlation across ability types	0.979**	0.972**	0.878**	0.807**
	(0.002)	(0.002)	(0.006)	(0.007)

This table shows the standard deviation of teacher value-added by student group, grade, and subject. We also show the correlation of value-added measures across student types. The estimation procedure for these standard deviations and correlations is described in Section 2.2. Standard errors are shown in parentheses. The significance of the correlation is tested against a null hypothesis of 1.
* Significance at the 10 percent level.
** Significance at the 5 percent level.

high-ability students lie between 0.8 and 0.9, representing a more serious rejection of the weak locational model in favor of a non-locational model.

Collectively, this evidence suggests that standard value-added models which rank teachers according to a single index of ability rely on theoretical foundations which are empirically rejected in the case of subject-heterogeneity models but are more reasonable for type-heterogeneity models.

4.5. The magnitude of value-added ranking error in a non-locational environment

While we formally reject the assumptions underlying the SLM and WLM, it is important to understand the degree to which value-added rankings of teachers are actually misleading. Consequently, we conclude this section with a couple of simulations of these effects. Essentially, we are calibrating our teacher ranking model using the estimated moments of the teacher value-added distribution. For simplicity we assume that there are either two subjects or two student types, 1 and 2, and that a teacher's joint ability to increase the achievement of these students, $\theta^1 = \mu^1\nu^1$ and $\theta^2 = \mu^2\nu^2$, follows a bivariate normal distribution. The means of θ^1 and θ^2 are normalized to zero. We select the remaining moments of the value-added distribution, σ^1, σ^2, and ρ, to reflect our empirical results in Tables 2 and 3. For each choice of parameters we simulate the experience of one million teachers, each with a draw from the bivariate normal distribution of θ^1 and θ^2. Since our earlier tables present estimates of the true parameters of the value-added distribution, these simulations explicitly abstract from estimation errors arising from either small student samples or noisy test measures and instead focus on the usefulness of value-added in the limit as the number of students assigned to each teacher approaches infinity.

Even under a non-locational model there are some pairs of teachers that can be strictly ranked by value-added. For any two teachers A and B, it may be that teacher A is strictly better than teacher B across both subjects or student types. A non-locational environment simply means that this is not true for *all* pairs of teachers. Indeed, the degree to which the weak locational model fails can be measured by the fraction of teacher pairs that do not display a dominant relationship. As this fraction approaches zero, the non-locational model approaches the weak locational model. This pairwise dominance relationship is also empirically important as it allows us to quantify one potential consequence of using standard value-added models in non-locational environments.

We begin in Table 4 by calculating the fraction of pairwise teacher comparisons in our simulated data that result in a definitive ranking. In the non-locational model, a definitive ranking between two teachers requires that the pairwise comparison yield a dominant relationship, that is one teacher must exceed the other in value-added across both subjects or types of student. Under the bivariate normal assumption for the joint value-added distribution, the result depends only on the correlation between θ^1 and θ^2. When ρ is 0.6, approximately 70 percent of pairwise comparisons of teachers yield dominant relationships. This fraction rises at an increasing rate with ρ. Because the $\rho = 1$ state corresponds to a locational model, a complete teacher ranking holds at that level. Even at a fairly high correlation level such as $\rho = 0.7$, a level consistent with the actual correlation of teacher value-added across subjects, approximately 25 percent of pairwise comparisons fail to identify a dominant relationship. When $\rho = 0.8$, the cross-type correlation in reading scores suggested by our data, more than 20 percent of comparisons cannot establish a dominant relationship. In such cases, without knowledge of the weights in the social welfare function, a substantial proportion of teachers cannot conceptually be ranked. Forcing a value-added ranking on these teachers will result in ranking errors.

Table 4
Fraction of pairwise comparisons that yield a dominant relationship.

ρ	Fraction
0.6	0.705
0.7	0.747
0.8	0.795
0.9	0.857
1	1.0

Each result is calculated from one million draws with the indicated parameters.

The problem becomes more pronounced when examining the dominance relationships between groups of teachers in a more narrow range of the quality distribution. After all, the marginal teachers in any firing decision will all be near the bottom of the value-added distribution. Since these teachers are closer in terms of weighted value-added, they tend to be closer in terms of value-added for each subject. This should lead to a lower fraction of strictly dominant relationships between teachers at the policy relevant margins.[9]

Table 5 examines the fraction of pairwise comparisons that yield dominant relationships as we compare teachers with varying degrees of similarity according to measured value-added. When comparing teachers within the same decile, using the correlation level of 0.8 found in cross-type reading scores, only 26 percent of pairwise comparisons yield a dominant relationship. That percentage increases to about 44 percent when comparing teachers in the same or adjacent deciles and 57 percent when looking at teachers less than or equal to two deciles apart. Thus these value-added measures seem to do a poor job of discerning between teachers who are relatively close in a distributional sense. Thus, these simulation results suggest that the precise identity of the teachers marked by value-added for sanctions (or rewards) is highly sensitive to the subject weights (or student–teacher assignment function) implicit in the standard value-added model. It is also worth noting, however, that value-added measures are rarely misleading in ranking teachers far apart in the observed distribution, providing the correct ranking 97 percent of the time when the teachers are at least three deciles apart.

[9] Similar problems arise when considering incentive policies that reward teachers at the top of the value-added distribution.

Table 5
The fraction of pairwise comparisons that yield dominant relationships.

ρ	All comparisons	Difference in decile						
		0	≤1	≤2	≤3	≥1	≥2	≥3
0.7	0.747	0.216	0.375	0.505	0.585	0.806	0.892	0.941
0.8	0.795	0.257	0.441	0.572	0.656	0.855	0.933	0.97
0.9	0.857	0.338	0.553	0.682	0.754	0.914	0.975	0.994
1.0	1	1	1	1	1	1	1	1

Each result is calculated from one million draws with the indicated parameters.

5. Welfare and policy implications

5.1. Welfare consequences of teacher firing

In the previous section, we demonstrated that statistical tests suggested by our model rejected the teacher subject-homogeneity assumption that underlies the use of standard value-added models. This implies that value-added models make mistakes in ranking teachers. We also provided some estimates of the fraction of mistakes that would result from using standard value-added rankings. However, when considering the welfare consequences (i.e. to student test scores) of using value-added to make school policy decisions it is also important to think about the magnitude of the mistakes, not just their number. In this section we use our previously introduced theoretical framework to derive conditions under which using value-added based rankings to make personnel decisions will improve student test scores. We then provide some empirical guidance about the possible magnitude of such improvement.

We begin with the observation that social preferences over student outcomes induce an ordering over collections of teachers, since two faculties can be compared by the social value of student outcomes that they induce. Social preferences over student outcomes define a marginal rate of transformation between the achievement of students in each subject through the welfare weights $\{M_{t(i)k}\}$. A measure of a teacher's value that is consistent with social welfare would consider this tradeoff between the achievement of students in each subject. The value-added measure of teacher performance, on the other hand, imposes an implicit trade-off defined by the subject weights N_m and N_r in subject heterogeneity models. Since this weight will in general differ from the socially optimal trade-off for almost all social welfare functions, in a non-locational setting value-added based policies may result in firing socially desirable teachers whose relative skills across subjects are more aligned with societal preferences but less aligned with the implicitly defined value-added trade-off. In the WLM environment, this does not occur because all teachers are at least weakly ordered and value-added correctly uncovers this ordering as the number of students becomes large.

In our framework, the social welfare induced by a particular faculty will depend on the qualities of the teachers in that faculty and on the way that teachers are assigned to students. Under the assumptions we have previously labeled the SLM and WLM, value-added measures of teacher quality can be used in personnel decisions to maximize student welfare in both type- and subject-heterogeneity models.

The following results for type and subject heterogeneity models demonstrate the conditions under which this occurs.

Proposition 1 (*Type-heterogeneity models*).
Assume that the social weights placed on types are not equal to the value-added weights placed on types. Then

(1) *If the environment does not satisfy the WLM with universal assignment then replacing a teacher who has below-average value-added can reduce social welfare.*
(2) *When comparing two teachers as potential candidates for firing, the closer their measured value-added, the higher the probability of firing the teacher who provides more social value.*
(3) *If both the social weight placed on all types is positive, and the covariance between $Q_{jt}\mu_{jt}$ and $Q_{jt'}\mu_{jt'}$ is positive for all t and t′ then replacing a teacher with low value-added will increase the social value of expected student outcomes.*

Proposition 2 (*Subject-heterogeneity models*).
Assume that the social weights placed on subjects are not equal to the value-added weights placed on subjects. Then

(1) *If the environment does not satisfy the WLM then replacing a teacher who has below-average value-added can reduce social welfare.*
(2) *When comparing two teachers as potential candidates for firing, the closer their measured value-added, the higher the probability of firing the teacher who provides more social value.*
(3) *If both the social weight and the value-added weights placed on all subjects are positive then replacing a teacher with low value-added will increase expected social welfare.*

The first result in both of these propositions shows that in non-locational settings the use of value-added to make firing decisions can lead to decreases in social welfare. The possibility of reducing social welfare as a result of using value-added measures of teacher quality stems from the fact that measured value-added imprecisely corresponds with the social value that a teacher provides. If the environment satisfies the conditions given in Lemmas 1 and 2, using value-added measurements of teacher quality to make personnel decisions will always improve social welfare because the value-added measurements establish a teacher quality ordering that is consistent with social preferences.

Since our empirical tests of environments in type-heterogeneity models were largely consistent with the

SLM, Proposition 1 suggests that in practice value-added can be a useful tool for personnel decisions in type-heterogeneity models. However, since we reject the WLM in subject-heterogeneity models, we need to investigate further the welfare consequences of using value-added measurements to make policy decisions in subject-heterogeneity contexts when the WLM fails.

The second result demonstrates that the probability of mistakenly firing a teacher is larger if her value-added score is closer to the comparison group of teachers. That is, when comparing two teachers whose measured value-added is very close to the 10th percentile, the likelihood of firing the teacher with the higher social value is higher than when comparing a teacher in the 10th percentile to one in the 50th percentile.

However, the third result states that even though the value-added weights for each subject (or type) and the social weights may not be equal, the value-added measure of teacher quality still provides information about the quality of the teacher. On average, teachers with higher value-added measures will have higher social value-added. Hence, value-added does not need to be "correct" in order for it to be useful for increasing student achievement. If teacher performance is positively correlated across subjects and student types, even when value-added misranks teachers the difference in social value will typically be small.

In summary, the theory suggests that personnel policies based on value-added measures will improve student achievement under all but extremely unlikely circumstances, even when the assumption of teacher homogeneity is violated. In particular, when answering the implicit question of how severe an assumption violation is too severe, our theory says too severe starts when the correlation of ability across subjects becomes negative. Intuitively, as long as the correlation is positive, in the absence of social preferences that ignore a particular subject, value-added is useful because it provides information, even if the information is noisy. Similarly, though we concentrate on personnel policies that target the lower tail of the teacher quality distribution, our results also imply that using value-added to reward teachers at the top of the distribution would also lead to positive student welfare gains so long as it actually increases the retention rate of high-quality teachers.

Beyond this theoretical point, it is also desirable to have some sense of how large the test score improvement would be. One common policy proposal is to use value-added in making determinations of which teachers to fire (Gordon et al., 2006). For concreteness, we simulate a policy in which the teachers in the bottom 10 percent of the measured teacher value-added distribution are fired and replaced with teachers that have the same average quality as the original distribution. There is nothing special about the 10 percent figure, nor are we suggesting that it is a likely policy outcome. We picked 10 percent because we feel that larger interventions may be politically infeasible and because simulating larger interventions run an increasing risk of ignoring progressively larger supply-curve effects.

The student achievement effects of this simulation exercise can be found in the first two columns of Table 6.

Table 6
Student achievement gains from firing teachers at the bottom of the value-added distribution.

ρ	Replace bottom 10%		Replace bottom 5%		Replace bottom 2%	
	Math	Reading	Math	Reading	Math	Reading
0.7	0.030	0.015	0.017	0.009	0.008	0.004
0.8	0.030	0.016	0.018	0.009	0.008	0.004
0.9	0.030	0.016	0.018	0.01	0.008	0.004
1	0.031	0.017	0.018	0.01	0.008	0.005

Each result is calculated from one million draws with the indicated parameters.

We find that when the correlation of teacher ability across subjects is 0.7, as in our data, students do indeed benefit, with achievement performance increasing by about 0.015 standard deviations in reading and 0.030 standard deviations in math. Indeed the table suggests that similar gains would be found at higher correlations as well. The remaining columns show that interventions that effect a smaller percentage of the teacher labor force will create smaller positive achievement effects, although the effect diminishes at a slower rate than a linear projection would suggest. Thus replacing the bottom two percent of teacher with average teachers would still increase student achievement by 0.008 standard deviations in math and 0.004 standard deviations in reading.

As suggested in our formal discussion, a policy that uses value-added measures to fire teachers performs relatively well, even though value-added provides a theoretically unfounded ranking, because the positive (but not perfect) teacher ability correlations across subjects result in a relatively small difference in ability between misranked teachers. Of course, just because replacing a low value-added teacher with an average value-added teacher raises both math and reading achievement on average does not mean that there is not a socially superior firing policy based on a different teacher performance metric. For example, if society values reading more on the margin than math, basing personnel policies an alternative teacher performance metric may bring about higher welfare increases if the reading improvements rise at the expense of smaller improvements in mathematics.

As a reminder of the degree to which teachers are misranked in the neighborhood of a firing policy cutoff, we briefly return to the results of Table 5. This table, given the empirically observed cross-subject correlation of teacher ability, suggests that only 22 percent of pairwise comparisons within a value-added decile yield dominant relationships. Consequently, the decision as to whether one teacher is better than another from nearby in the teacher ability distribution will often depend on how subjects (or student types) are weighted. If the value-added weights differ systematically from the social weights, there is a large scope for misrankings.

Furthermore, the magnitude of the achievement gains in Table 6 highlight the degree to which advocates of firing policies depend on its fundamentally changing the nature of the teaching talent distribution in the long run to produce anything more than modest results. Alternatively, they may depend on multiplying such results across

multiple years with no achievement decay. In contrast, the lack of clear, discernable dominant relationships near a firing policy cutoff based on value-added metrics may lead teachers and policymakers to view such policies as unfair and capricious. This may affect the ability to improve the general quality of the potential teacher pool in the long run.

5.2. Specialization policies

A conceptual problem with the most commonly suggested policy uses of value-added test scores, using teacher rankings for merit pay or teacher retention decisions, is that they fail to provide concrete ways for current teachers to become better. This stems from the failure of teacher value-added rankings to answer the question of why a particular teacher is better than another. Such policies also tend to rely on fundamentally speculative long-run changes in the teacher labor pool to produce results. We would argue that this policy mindset results from the almost complete focus on cross-teacher heterogeneity to the exclusion of within-teacher heterogeneity. While the former may be far greater in magnitude, it is hard to use as the basis of successful policy changes in the short or medium-run.

In contrast, it may be possible for schools to use value-added test score measures to leverage within-teacher heterogeneity to improve the productivity of their current teacher workforces by assigning teachers to students or subjects in which they enjoy a teaching comparative advantage. Indeed, our model suggests very specific conditions under which this will work.

Proposition 3. *For both type- and subject-heterogeneity models, teacher specialization will be optimal and non-specialization will be generically sub-optimal unless the SLM holds and the social weight put on either types or subjects is equal.*

This proposition means that teacher assignment policies that are common across teachers, which could include random or balanced assignment across student types or subjects may be desirable for econometricians, but will generally not maximize welfare. Instead, social welfare can be increased by assigning teachers to student types or subjects with which they match well. In many cases, the expected achievement of all student types and subjects can increase as a result.

In Table 7, we attempt to quantify the student achievement benefits that could accrue to greater teacher subject specialization in grades 4–5. In this simulation we assign each teacher exclusively to the subject in which they have demonstrated the highest value-added (normalized by the standard deviation in value-added in each subject). Hence, a teacher is assigned to reading if $\frac{\theta^r}{\sigma^r} > \frac{\theta^m}{\sigma^m}$. The results suggest that this produces gains in average student test scores in both subjects, which increase as the cross subject correlation of teacher value-added drops. Furthermore, for the values of ρ we estimated in our data, the subject specialization of teachers produces gains at least fifty percent larger than a ten percent teacher firing policy. In fact to get the same gains from a teacher firing policy, our data suggest you would have to be able to replace approximately the bottom value-added quartile of current teachers with average teachers.

Table 7
Student achievement gains from teacher subject specialization.

ρ	Math	Reading
0.7	0.054	0.030
0.8	0.044	0.025
0.9	0.031	0.018
1	0.000	0.000

Each result is calculated from one million draws with the indicated parameters.

Table 8
Student reading achievement gains from sorting students to teachers with comparative advantage in teaching high- or low-ability students.

ρ	Reading
0.7	0.030
0.8	0.025
0.9	0.017
1	0.000

Each result is calculated from one million draws with the indicated parameters.

The table also provides suggestive evidence about how the lack of a theoretical framework inhibits policy formation. If we start by assuming that we can definitively rank all teachers, we place ourselves in the strong locational model. Here, there is no need to consider cross subject correlations since they will be perfect. On the other hand, an explicit framework that considers how such rankings might be conceptually problematic leads to the insight that within-teacher heterogeneity might provide useful information.

We next consider the potential benefits of directed matching between teachers and students of different types. Like subject specialization, the potential for teacher, as opposed to peer, effects in matching is generally ignored. In fact, such matching stands in practical opposition to the measurement ideal of as good as random assignment, and is often treated as an explicit negative in empirical studies. Our results from Table 3, however, suggest that while the ability to teach math is almost perfectly correlated across high and low initial ability students, there may be substantial differences in a teacher's ability to teach reading to different types of students.[10] In Table 8, we examine the potential improvements in student reading achievement that could be realized in our simulated environment from assigning teachers to students according to the teachers' comparative advantage and the students' type. More specifically, if $\frac{\theta^1}{\sigma^1} > \frac{\theta^2}{\sigma^2}$ we assign a teacher all type 1 (high initial ability) children. If the opposite result holds, we assign a teacher all low initial ability children. Because we have implicitly

[10] It seems that such student sorting is uncommon in our data. Regression results imply that teachers with a comparative advantage in teaching high-ability reading students do not have a higher proportion of high-ability reading students assigned to them.

assumed equal numbers of high-ability and low-ability children, this policy allows all students to be served on average.

As expected from our theoretical development and earlier results on subject specialization, the table shows that this policy provides a Pareto improvement relative to the random assignment of teachers and students. More interestingly, the benefits from matching by student type again exceed the benefits of the alternative policy of firing 10 percent of teachers, even when the correlation of teacher ability across types of students is as high as.9. In fact, with this value of ρ, matching raises student academic achievement by as much as 0.02 standard deviations for both high- and low-ability students. For lower values of ρ, the benefits of matching are even larger.

The simulations in this paper present relatively abstract scenarios. In particular, they compare policies that realize the entire potential gains from either firing or matching where in reality those potential gains should be thought of as bounds that an actual policy would be unlikely to attain. Of course, the theoretical models in this paper depend on asymptotical inference properties, whereas actual value-added measurements may be based on small samples. The empirical estimates of this paper are also calculated to produce estimates of moments of the value-added distribution purged of testing noise. Actual implementations of value-added ratings may fail to achieve either of these ideals and may realize only a fraction of the theoretically posited gains.

However, beyond this standard caveat there are other, non-sampling issues. For example, there is a potential concern that the smaller number of teachers within a school or grade level might allow for less specialization than possible in our simulations. To gauge the extent of this problem, we perform an additional simulation of the effects of teacher subject specialization under the constraint that all school faculties contain only two randomly drawn teachers per grade. Even in this worst case specialization scenario, more than two-thirds of the unrestricted potential student test score gains are still realized. [11] Also, there may be complicating effects that arise from younger students having to adjust to a classroom format with more teachers in a given school day. Finally, just as students may take unanticipated actions to undermine policy attempts to create optimal peer matches (Carrell, Sacerdote, & West, 2013), either students or teachers may take unanticipated actions that undermine attempts at better student–teacher matching.

In evaluating value-added for personnel purposes, we have actually maintained assumptions that are favorable toward a teacher firing policy. For example, we have heretofore ignored the evidence that teacher value-added measures based on standardized tests have only moderate correlations (0.3–0.5) with more demanding open response performance measures (Rothstein, 2011). Similarly, we have elided the potential of teachers to improve their value added in response to specific interventions, in the manner suggested by Taylor and Tyler (2009). It could also be the case that it is unrealistic, in either a political economy or supply elasticity sense, to fire so large a group of teachers (ten percent). Even if it were possible to do so in the short run it could serve to discourage potential teachers, especially those with the best outside alternatives, from entering the profession in the future. Finally, our assumption that replacement teachers would be on average, as good as existing ones is only possible if districts are currently unable to make any quality-based decisions. There is certainly empirical evidence to support the idea that the potential pool of teachers is of lower average quality than current teachers (Jepsen & Rivkin, 2009). In any of the above cases our simulations would overestimate the realized social benefits of a teacher firing policy. For example, in our simulations, replacing the bottom ten percent of current teachers with teachers in the fortieth percentile of value-added would produce student test gains of 0.26 standard deviations in math (rather than the 0.30 in our tables), and 0.13 in reading (as opposed to the reported 0.16). Conversely, our comparison could understate the possible student gains from value-added based firing as it assumes the firings have no effect on the performance of remaining teachers. This is partly due to a lack of good evidence as to the extent such a policy would have, "pour encourager les autres."

It is also possible that these simulations understate the benefits of teacher specialization. In particular, greater specialization may lead to more focused specific human capital investment. Also, each teacher would spend a higher fraction of time in a single task allowing them to reap the benefits of experience and learning by doing in a shorter period of calendar time.[12]

6. Conclusion

Despite the clear policy relevance of teacher value-added in current educational debates, economists have been slow to develop explicit formalizations that allow us to weigh the cost–benefit tradeoffs that occur when the assumptions of value-added are violated. Consequently, these tradeoffs are not clearly understood by those debating policy. In this paper we present a model that helps clarify how heterogeneous teacher effects across subjects and differentiated student types would affect our thinking about the use of value-added models for teacher evaluation. In particular we show how value-added measurements behave in three potential environments, and derive model predictions to serve as empirical tests in distinguishing among the environments. In two of these

[11] The simulation assumes a cross-subject value-added correlation of 0.7 and produces math gains from students of 0.038 and reading gains of 0.021.

[12] This study implicitly assumes teacher time use-effects are subject neutral. That is, it is possible that part of the measured subject differences in teacher ability really come from the fact that teachers spend more class time on subjects they are a little better at. This would create a downward bias in our measure of ability correlations. Of course the opposite is also possible, namely teachers spend more class time on subjects they feel they are worse at teaching in order to partially compensate for this ability difference. This would imply our estimate of cross-subject ability correlations is too high. Our data do not provide us with any information on the subject, however.

cases, which we call the strong and weak locational models, value-added can provide a definitive teacher ranking, though the latter case may require specific restrictions on how teachers are assigned to students. Using matched teacher–student longitudinal data from North Carolina, our tests show that the non-locational environment is the best description for our data, particularly for subject heterogeneity models.

Although our empirical tests show that the teacher heterogeneity assumption does not hold, we use our model to show that under any reasonable conditions, commonly suggested personnel policies based on value-added will raise student achievement. We then use estimated moments of the teacher value-added distribution to simulate the actual student achievement effects of proceeding with a value-added based teacher firing policy. As predicted, we find that this policy improves student achievement, though the positive student test score effects are small in magnitude. Thus, with any social welfare function that does not weight fairness to teachers (or weights it little) and does not allow for adjustment on teacher assignment margins, the policy might be close to optimal.

We also provide simulation results that show the degree to which value-added measures provide a systematic misranking of teachers when the teacher heterogeneity assumption is violated. We find a large number of cases in which such misrankings would result in a teacher firing policy based on value-added measures firing the wrong teachers. While these misrankings do not result in lower student welfare, they are certainly unfair and arbitrary to teachers whose abilities put them near the policy cutoff. Such perceptions of unfairness may also affect the long-run ability to attract better quality teachers into the profession.

Thus, given the heterogeneity of teacher ability across subjects and types of students, we suggest a potential alternative policy use of value-added information. In particular, we show that the information from value-added measures might be better employed to improve student achievement using the current teacher labor force by leveraging information about differential teacher ability across subjects and student types. Our further simulations show that a policy of teacher specialization in their best subject or exploiting teacher–student match quality produces greater achievement gains than the firing policy. In fact, the simulation numbers indicate that a personnel policy would have to substitute the bottom 25 percent of teachers with average replacements to improve student math scores by the same amount as a teacher specialization policy. To get equivalent reading test gains to the subject specialization policy, the firing policy would have to replace the bottom 30 percent. Furthermore, while the one-year effects of a specialization policy may seem small, if the effects persist and accumulate across all elementary grades, they would provide math achievement gains similar to the one-year effect of providing a student with a two-standard deviation better value-added teacher.

This paper highlights the importance of a theoretical framework in making policy decisions. It provides such a framework for thinking about the possible uses of teacher rankings, as well as an explicit evaluation of the welfare implications of the inability to correctly rank teachers. It also highlights the underappreciated role of within-teacher ability heterogeneity. More generally, it informs the literature on the use of competency tests to rank employees when their skills may be heterogeneous across tasks. Our results suggest that employers might realize greater gains by increasing the specialization of their employees' tasks rather than attempting to replace them with hypothetically better employees.

Appendix A. Proofs

Proof of Lemmas 1, 2 and 3

We start by proving lemmas 2 and 3 and then show how Lemma 1 is a special case of these.

In type-heterogeneity models, under the SLM $\nu_t = \nu_{t'}$ for all t, t' and for any teacher j and types t and t', $\mu_{jt} = \mu_{jt'}$. Let ν be this common student sensitivity to teacher input and let μ_j be the teacher's input for each student type. By definition $j \gg j'$ if $\mu_j > \mu_{j'}$ which implies that $\mu_j \nu > \mu_{j'} \nu$. Under these assumptions

$$EV_j = E\left[\frac{1}{I_j}\sum_{i=1}^{I_j} A_{t(i)}\right] = \frac{1}{I_j}\sum_{i=1}^{I_j} EA_{t(i)} = \frac{1}{I_j}\sum_{i=1}^{I_j}\sum_{t=1}^{T} Q_{jt}\mu_{jt}\nu_t = \frac{1}{I_j}\sum_{i=1}^{I_j}\mu_i\nu\sum_{t=1}^{T} Q_{jt} = \mu_j\nu \tag{A.1}$$

Under the assumption that $Var(\epsilon_i)$ are uniformly bounded and that $Cov(\epsilon_i, \epsilon_{i'}) = 0$ for all i, i', the strong law of large numbers implies that $V_j \to \mu_j\nu$ almost surely.[13] The result then follows immediately.

To see that Lemma 3 holds, recall that we have assumed that all teachers have equal class sizes so although the set of students assigned to j and j' are different, their total numbers I_j and $I_{j'}$ are equal. Thus, if $Q_j \equiv Q_{j'}$ then

$$EV_j - EV_{j'} = \frac{1}{I_j}\sum_{i=1}^{I_j}\sum_{t=1}^{T} Q_{jt}\mu_{jt}\nu_t - \frac{1}{I_j}\sum_{i=1}^{I_{j'}}\sum_{t=1}^{T} Q_{jt}\mu_{jt}\nu_t = \frac{1}{S}\sum_{s=1}^{S}\sum_{t=1}^{T} Q_{st}(\mu_{jt} - \mu_{j't})\nu_t \geq 0. \tag{A.2}$$

As long as $Q_{jt} > 0$ for at least one t for which $\mu_{jt} > \mu_{j't}$ then the above difference will be bounded above zero and so by the strong law of large numbers

$$P\left\{\lim_{I_j\to\infty} V_j > \lim_{I_{j'}\to\infty} V_{j'}\right\} = 1 \tag{A.3}$$

To see the second result define $t_{j'} = \arg\max_t \mu_{j't}$ and $t_j = \arg\min_t \mu_{jt}$. Letting $Q_{jt_j} = 1 = Q_{j't_{j'}}$ then

$$EV_j - EV_{j'} = \frac{1}{S}\sum_{s=1}^{S}\sum_{t=1}^{T} Q_{it}(\mu_{j't})\nu_t < 0, \tag{A.4}$$

[13] See Chung (2001) Theorem 5.1.2 for a statement and proof of this result.

which implies that

$$P\left\{\lim_{I_{j'}\to\infty} V_{j'} > \lim_{I_j\to\infty} V_j\right\} = 1. \tag{A.5}$$

Thus, under the assignment function defined, the value-added estimator $V(\cdot)$ will misorder teachers j and j', even if there is an unlimited amount of data on the value-added of each teacher type.

While this is an extreme example, the continuity of $V_j(S)$ in Q reveals that for any probability weighting function sufficiently close to the one just described the result will also hold.

Finally, we note that subject heterogeneity models are analogous to type-heterogeneity models in the sense that the weights (N_m, N_r) play the role of assignment functions. Since these assignment functions are the same across all teachers, the argument in the proof of Lemma 3 can be applied directly to prove Lemma 1. □

Proof of Implication 1

Let $\bar{V}_j(u)$ be the expected value-added estimate across teachers using students of group u. By definition

$$\hat{\sigma}^2(V_j(u)) = \frac{1}{J}\sum_{j=1}^{J}\left(V_j(u) - \bar{V}_j(u)\right)^2 \tag{A.6}$$

Recall that the value-added estimate $V_j(u)$ is defined as

$$V_j(u) = \frac{1}{I_{ju}}\sum_{i=1}^{I_{ju}}(\mu_j\nu_u + \epsilon_i). \tag{A.7}$$

Here the teacher input μ_j is not indexed by group because under the WLM and SLM this input is common across student groups. Define $\bar{\epsilon}^u_{ij} = \frac{1}{I_{ju}}\sum_{i=1}^{I_{ju}}\epsilon_i$ and $\bar{\epsilon}^u = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{I_{ju}}\sum_{i=1}^{I_{ju}}\epsilon_i$. Substituting (A.7) into (A.6) simplifying yields

$$\hat{\sigma}^2(V_j(u)) = \nu_u^2\sigma^2(\mu_j) + \sigma^2(\bar{\epsilon}^u_{ij}) + \frac{2}{J}\nu_u\sum_{j=1}^{J}(\bar{\epsilon}^u_{ij} - \bar{\epsilon}^u)(\mu_j - \bar{\mu}) \tag{A.8}$$

For each teacher j, as $I_{ju}\to\infty$, both $\bar{\epsilon}^u_{ij}$ and $\bar{\epsilon}^u$ converge to zero almost surely which implies that $\hat{\sigma}^2(V_j(u)) \to \nu_u^2\sigma^2(\mu_j)$ almost surely.

The covariance between value-added across types can be calculated as

$$Cov(V_j(u), V_j(u')) = E(\mu_{ju}\nu_u + \bar{\epsilon}^u_{ij} - \overline{\mu_{ju}}\nu_u)(\mu_{ju'}\nu_{u'} + \bar{\epsilon}^u_{ij'} - \overline{\mu_{ju'}}\nu_{u'}).$$

As I_{ju} increases without bound, this converges almost surely to $\nu_u\nu_{u'}\sigma^2(\mu)$. This fact, combined with the previous result shows that

$$P\left\{\lim_{I_{ju}\to\infty}\rho(V_u, V_{u'}) = \frac{Cov(V_u, V_{u'})}{\sigma(V_u)\sigma(V_{u'})} = 1\right\} = 1 \tag{A.10}$$

□

Proof and discussion of Propositions 1 and 2

We will couch the proof of this proposition in the type-heterogeneity model. Discussion of how to alter the proof for subject-heterogeneity models is made at the end.

Consider a set of assignment functions $\{Q_j\}_{j\in\mathcal{J}}$, one for each teacher. Assume that classroom sizes are the same across teachers so each teacher has I/J students. Further assume that the fraction of all students that are of type t is ω_t. The requirement that each student have a teacher implies that

$$\sum_{j\in\mathcal{J}}\frac{I}{J}Q_{jt} = I\omega_t \tag{A.11}$$

Calculation shows that the fraction of students of type t assigned to teacher j is

$$P[j|t] = \frac{P[t|j]P[j]}{P[t]} = \frac{Q_{jt}}{J\omega_t} \tag{A.12}$$

Given this set of assignment functions, the expected achievement for a student of type t is

$$E[A(t)|\{Q_j\}_{j\in\mathcal{J}}] = \sum_j P[j|t]E(A|j) = \sum_{j\in\mathcal{J}} P[j|t]\mu_{jt}\nu_t = \frac{1}{J\omega_t}\sum_j Q_{jt}\mu_{jt}\nu_t \tag{A.13}$$

Therefore, the

$$U(A) = \sum_{i=1}^{I} M_{t(i)}EA_i = I\sum_{t=1}^{T}\omega(t)M_tE[A(t)|Q] = I\sum_{t\in\mathcal{T}}\omega_tM_t\frac{1}{\omega(t)}\sum_{j\in\mathcal{J}}Q_j(t)\mu_{jt}\nu_t = I\sum_{t\in\mathcal{T}}M_t\sum_{j\in\mathcal{J}}Q_j(t)\mu_{jt}\nu_t \tag{A.14}$$

When the environment does not satisfy the strict locational model, the use of teacher value-added in making hiring and firing decisions may lead to the formation of a faculty that is less socially desirable than the status quo.

To see this, assume that the WLM is incorrectly assumed to hold and that the teacher for whom the value-added estimator V is lowest is fired. To see whether this move was desirable it is necessary to compare social welfare under the previous faculty and the new faculty.

Under our assumption of linear social welfare, the difference between social welfare under the two regimes is

$$I\sum_{t\in\mathcal{T}}M_t\sum_{j\in\mathcal{J}}Q_j(t)\mu_{jt} - I\sum_{t\in\mathcal{T}}M_t\sum_{j\in\mathcal{J}'}Q_j(t)\mu_{jt} \tag{A.15}$$

Assume that the teacher that was fired is labeled 1 and the draw from the distribution of potential alternative teachers is 1′. Furthermore, assume that the assignment function Q does not change. Then the difference above becomes

$$\sum_{t\in\mathcal{T}}M_tQ_1(t)(\mu_{1t} - \mu_{1't})\nu_t \tag{A.16}$$

Under value-added, the teacher was fired because

$$V_1 - EV_{1'} < 0. \tag{A.17}$$

Under the (incorrect) assumption that the WLM holds, a teacher from outside the faculty can be characterized by a single parameter $\overline{\mu}$. Thus, using teacher value-added as a metric on which to base firing decisions implies that a teacher is fired if

$$\begin{aligned} EV_1 - EV_{1'} &= E\frac{1}{I_1}\sum_{i=1}^{I_1}(\mu_1 - \overline{\mu})\nu_t < 0 \\ &= \frac{1}{I_1}\sum_{i=1}^{I_1}\sum_{t\in T} Q_{1t}(\mu_{1t} - \overline{\mu})\nu_t \end{aligned} \tag{A.18}$$

A teacher that is fired under the incorrect WLM assumption but would not have been fired otherwise satisfies

$$\begin{aligned} \sum_{t\in\mathcal{T}} M_t Q_{1t}(\mu_{1t} - \mu_{1't})\nu_t &> 0 \\ \frac{1}{I_1}\sum_{i=1}^{I_1}(\mu_1 - \overline{\mu})\nu_t &< 0 \end{aligned} \tag{A.19}$$

On average (i.e. taking the expected value of the second expression), the teacher that will satisfy these conditions is one for which

$$\begin{aligned} \sum_{t\in\mathcal{T}} M_t Q_{1t}(\mu_{1t} - \mu_{1't})\nu_t &> 0 \\ \sum_{t\in\mathcal{T}} Q_{1t}(\mu_{1t} - \overline{\mu})\nu_t &< 0 \end{aligned} \tag{A.20}$$

These inequalities will be simultaneously satisfied for those teachers for which $\sum_t M_t Q_{1t}\mu_{1t}\nu_t$ is relatively large and $\sum_t Q_{1t}\mu_{jt}\nu_t$ is relatively small. The first sum will be large relative to the second sum when teacher 1 has a relative advantage at teaching students with high social value (M_t is large) and that have relatively high sensitivity to teacher inputs.

In subject-heterogeneity models, the matching function is replaced by the (common) term N_m or N_r representing the value-added weights for math and reading. It is clear, since these weights behave like a common assignment function that under the assumptions of the WLM, using value-added will correctly order teachers in terms of social value so replacing the worst teachers will increase social welfare. The problems that arise when the WLM fails are identical to those in type-heterogeneity models.

The proof of item two is as follows. First let $A(W)$ be the set of teachers $\{j'\}$ for whom $W_{j'} < W$ and let $B(V)$ be the set of teachers for whom the expected value of teacher value-added is at least V. Notice that the size of $A(W)$ is increasing in W and the size of $B(V)$ is decreasing in V. The set of teachers that have lower social value-added than teacher j but higher (expected value of) value-added is the set $A(W_j) \cap B(V_j)$. The first claim of this proposition follows directly from the fact that $B(V)$ is decreasing in V. Hence, for a teacher with value-added V_j, the likelihood that replacing teacher j with a teacher with value-added V will lead to declines in social value decreases as the value-added of the replacement increases.

We now show that the probability of lowering social value is positive as the replacement's value-added approaches that of teacher j. We will assume that $M_m/M_r > N_m/N_r$. For a hypothetical teacher j' for whom $V_{j'} = V_j$, if $(\mu_{j'm}, \mu_{j'r}) \neq (\mu_{jm}, \mu_{jr})$ then $W_{j'} < W_j$ if $\mu_{j'm}/\mu_{j'r} < \mu_{jm}/\mu_{jr}$. By the continuity of value-added and social value-added in (μ_{jm}, μ_{jr}), and since the value-added and social value-added weights are positive for all subjects, there then exists an $\epsilon > 0$ for which the set of teachers $\{j'\}$ for which and $W_{j'} < W_j$ has positive measure, which proves the result.

Finally, we discuss the proof of item three. We show this result first for subject heterogeneity models, as the extension to type heterogeneity models is straightforward. Let (μ_{jm}, μ_{jr}) be jointly normally distributed with mean (μ_m, μ_r) and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_m^2 & \sigma_{mr} \\ \sigma_{mr} & \sigma_r^2 \end{pmatrix}. \tag{A.21}$$

Let $W_j = M_m\mu_{jm}\nu_m + M_r\mu_{jr}\nu_r$ be the social value-added of teacher j. Social welfare is increasing in W_j. We are interested in calculating $E[W_j|V_j]$. Since both W_j and V_j are linear combinations of (μ_{jm}, μ_{jr}), they are jointly normally distributed. This implies that

$$E[W_j|V_j] = E[W_j] + \frac{\mathrm{cov}(W_j, V_j)}{\sigma_{V_j}}(V_j - E[V_j]) \tag{A.22}$$

Social welfare will on average increase by replacing teachers with below average value-added if the above expression is increasing in V_j. This will occur if and only if $\mathrm{cov}(W_j, V_j) > 0$. This term is

$$\begin{aligned} \mathrm{cov}(W_j, V_j) = E(&M_m\mu_{jm}\nu_m + M_r\mu_{jr}\nu_r - M_m\boldsymbol{\mu}_m\nu_m \\ &+ M_r\boldsymbol{\mu}_r\nu_r) \times (N_m\mu_{jm}\nu_m + N_r\mu_{jr}\nu_r \\ &- N_m\boldsymbol{\mu}_m\nu_m + N_r\boldsymbol{\mu}_r\nu_r + \overline{\epsilon}_j) \end{aligned} \tag{A.23}$$

where $\overline{\epsilon}_j$ is the weighted average of student level idiosyncratic deviations given in Eq. (4.1). Simplification of this expression shows that

$$\begin{aligned} \mathrm{cov}(W_j, V_j) = &M_m N_m \sigma_m^2 + (M_m N_r + M_r N_m)\sigma_{mr} \\ &+ M_r N_r \sigma_m^2. \end{aligned} \tag{A.24}$$

When M_m and M_r are positive and $N_a \in (0, 1)$, this number will be positive.

For type-heterogeneity models, the covariance between V_j and W_j is

$$\begin{aligned} \mathrm{cov}(W_j, V_j) &= E\left(\sum_{t=1}^{T} M_t Q_{jt}\mu_{jt}\nu_t - \sum_{t=1}^{T} EM_t Q_{jt}\mu_{jt}\nu_m\right) \\ &\times \left(\sum_{t=1}^{T} Q_{jt}\mu_{jt}\nu_t + \overline{\epsilon}_j - \sum_{t=1}^{T} EQ_{jt}\mu_{jt}\nu_m\right) \\ &= \sum_{t=1}^{T} M_t\sigma^2(Q_{jt}\mu_{jt}) + 2\sum_{t\neq t'} \nu_t\nu_{t'} EQ_{jt}Q_{jt'}(\mu_{jt} - \boldsymbol{\mu}_t) \\ &\times (\mu_{jt'} - \boldsymbol{\mu}_{t'}). \end{aligned}$$

The variance terms in this equation will be positive, so under the restriction that the covariance between $Q_{jt}\mu_{jt}$ and $Q_{jt'}\mu_{jt'}$ is positive then the value-added measure will provide

information on the social value-added of each teacher. This imposed condition says that teachers are not matched with students with whom they are less skilled so much that this matching undoes the positive correlation between μ_{jt} and $\mu_{jt'}$. □

Proof of Proposition 3

The model presented in the paper allows us to solve for the optimal matching functions of teachers to students. This optimal function will be the solution to the problem

$$\max_{Q_1,\ldots,Q_J} E\sum_{t=1}^{T}\left(M_t \sum_{\{i:t(i)=t\}} A_i\right) \quad \text{s.t.} \sum_{t=1}^{T} Q_{jt} = 1 \text{ for all } j \tag{A.26}$$

where M_t is the social value placed on the achievement of a student of type t.

Simplification of the objective function implies that

$$\begin{aligned} E\sum_{t=1}^{T}\left(M_t \sum_{\{i:t(i)=t\}} A_i\right) &= \sum_{t=1}^{T}\left(M_t \sum_{\{i:t(i)=t\}} EA_i\right) \\ &= \sum_{t=1}^{T}\left(M_t \sum_{\{i:t(i)=t\}} \sum_{j=1}^{J} P[j|t(i)]\mu_{jt}\nu_t\right) \\ &= \sum_{t=1}^{T}\left(M_t \sum_{\{i:t(i)=t\}} \frac{1}{J\omega_t} \sum_{j=1}^{J} Q_{jt}\mu_{jt}\nu_t\right) \\ &= \sum_{t=1}^{T}\left(M_t \sum_{\{i:t(i)=t\}} \frac{1}{J\omega_t} \sum_{j=1}^{J} Q_{jt}\mu_{jt}\nu_t\right) \\ &= \frac{I}{J}\sum_{t=1}^{T}\left(M_t\nu_t \sum_{j=1}^{J} Q_{jt}\mu_{jt}\right) \end{aligned} \tag{A.27}$$

where the third equality follows from Eq. (A.13).

This is a linear function in the choice variables Q_{jt}. It can be seen that the solution to this problem will be to assign the best teachers first to the student type for which $M_t\mu_{jt}\nu_t$ is largest. Then of those teachers that are left, it will then be optimal to assign teachers to students when $M_t\mu_{jt}\nu_t$ is second largest and so on.

For the set of linear social welfare functions, optimal matching, as defined by the set of assignment functions $\{Q_j\}_{j\in\mathcal{J}}$, is the solution to a linear programming problem.

Since the set of feasible matching functions is a compact set, the fundamental theorem of linear programming says that a solution to the problem exists and one such solution can be found at an extreme point of the set of feasible matching functions. Furthermore, if there exist two points $\{Q_j\}_{j\in\mathcal{J}}$ and $\{Q'_j\}_{j\in\mathcal{J}}$ in the relative interior of the set of feasible matching functions for which $EU(\cdot,\{Q_j\}_{j\in\mathcal{J}}) \neq EU(\cdot,\{Q'_j\}_{j\in\mathcal{J}})$ then there is no solution to this program that is in the relative interior of the feasible set.

It is straightforward to show that if the WLM fails to hold then a common assignment function (which is in the relative interior of the feasible set of assignment functions) can be shown to provide local perturbations in the assignment function that alter social welfare. Hence, there is no optimal matching function in the relative interior of the feasible set which implies specifically that a common matching function is not socially optimal. □

References

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

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavorial Statistics, 29*(1), 37–66.

Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica, 81*(3), 855–882.

Chung, K. L. (2001). *A course in probability theory*. San Diego: Academic Press.

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

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement?* NBER Working Paper Series (12828).

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job*. Brookings Institute Discussion Papers.

Guarino, C., Reckase, M., & Wooldridge, J. (2011). *Can value-added measures of teacher performance be trusted?* Education Policy Center Working Paper #18.

Hanushek, E., & Rivkin, S. (2004). *How to improve the supply of high quality teachers*. Brookings Papers on Education Policy.

Hanushek, E. A. (2002). *Creating a New Teaching Professionchap. Teacher deselection*. Urban Institute Press.

Harris, D. N. (2009). Would accountability based on teacher value-added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*, 319–350.

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

Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources, 44*(1), 223–250.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: And experimental evaluation*. NBER Working Paper Series: #14607.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. National Center on Performance Initiatives Working Paper #2007-03.

Lefgren, L., & Sims, D. P. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis, 34*(1), 109–121.

Lockwood, J. R., & McCaffrey, D. (2009). Exploring student–teacher interactions in longitudinal achievement data. *Education Finance and Policy, 4*(4), 439–467.

Monk, D. H. (1987). Assigning elementary pupils to their teachers. *Elementary School Journal, 88*(2), 166–187.

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

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

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

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

Rothstein, J. (2011). *Review of 'Learning about teaching: Initial findings from the Measures of Effective Teaching project'*. National Education Policy Center.

Springer, M., Ballou, D., Hamilton, L., Le, V., Lockwood, J., McCaffrey, D., Pepper, M., & Stecher, B. (2010). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching*. National Center on Performance Incentives at Vanderbilt University.

Suzumura, K. (2002). *Handbook of social choice and welfarechap. 1. Handbooks in economics*. North Holland.

Taylor, E. S., & Tyler, J. H. (2009). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628–3651.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*, 3–33.