TEACHER VALUE-ADDED IN THE ABSENCE OF ANNUAL TEST SCORES: UTILIZING TEACHER NETWORKS

Latest Version

Desislava Tartova *

February 24, 2025

Abstract

This paper proposes a novel methodology for estimating teacher value-added in the presence of non-random teacher-student sorting and the absence of annual standardized student test scores. Rather than relying on lagged student test scores to control for nonrandom teacher-student sorting on student ability, I exploit within-student across-subject variation in test scores and teacher "networks"—teachers in the same subject who teach groups of students observed with a unique teacher in another subject. The resulting estimates closely recover the true parameters of teacher value-added in Monte Carlo simulations and align well with estimates from standard methodologies in New York City data where lagged test scores are available. The methodology substantially expands the research on teacher value-added, as the majority of educational settings do not rely on standardized testing in consecutive grade levels. I apply the method to French middle school teachers and find that a 1 SD increase in teacher value-added within school improves student scores by 0.10 SD in Math and 0.07 SD in French. I show that using a "hybrid" methodology—which augments the network estimator to control for lagged scores— in settings where lagged scores are available can outperform standard methods under specific sorting patterns by accounting for sorting on time-varying student unobservables.

*Paris School of Economics (PSE). Email address: desislava.tartova@psemail.eu. I am very grateful to Luc Behaghel, Julien Grenet and Jonah Rockoff for their guidance and support. I also thank Natalie Bau, Barbara Biasi, Sandra Black, Nina Guyon, Antoine Hubert de Fraisse, Adrien Matray, Sebastián Otero, Thomas Piketty, Camille Terrier, Yves Zenou, seminar participants at IZA, CESifo/ifo, EALE, EEA, Columbia University, SSE, ENS, PSE, and Paris 1 Sorbonne. I am grateful to the Direction de l'Évaluation, de la Prospective et de la Performance (DEPP) at the French Ministry of Education for granting me access to the data used in this study. I am also grateful to Jonah Rockoff for providing me with the dataset used in Chetty, Friedman, and Rockoff (2014a) to complete this analysis.

1 Introduction

Teacher value-added (hereafter TVA)—the causal effect of a teacher on their students' test scores—has been widely recognized as a key determinant of both academic performance (Rockoff, 2004, Kane & Staiger, 2008, Chetty et al., 2014a) and long-term socioeconomic outcomes, such as college attendance and earnings (Chetty, Friedman, & Rockoff, 2014b). TVA has thus become central to education research and policymaking, with several U.S. states incorporating TVA estimates into teacher evaluation and compensation systems. This has spurred an extensive literature on TVA estimation, both to assess its validity for policy decisions[1] and to explore broader questions about teacher effectiveness.[2]

Yet, TVA estimation has remained a largely U.S.-centric debate, as existing methods rely on annual standardized testing, a feature uncommon in most education systems. In the U.S., middle school students take standardized exams every year, allowing researchers to control for unobservable student traits—such as ability or motivation—using prior test scores. This is crucial when teacher-student assignment is nonrandom—for instance, if better teachers are systematically sorted with better students. Without lagged scores, TVA estimates may be subject to omitted variable bias. Outside of the U.S., standardized testing is far less frequent. For example, Denmark is the only European country to administer annual middle school exams, while most countries test students only at key educational stages. This pattern limits the applicability of existing TVA methods and underscores the need for approaches that do not rely on consecutive test scores.

This paper develops a methodology for estimating plausibly unbiased TVA in settings where students do not sit standardized tests yearly. Instead of relying on lagged test scores to control for student ability, I exploit within-student, across-subject variation and "networks" of teachers—teachers in the same subject who teach sets of students observed with a unique teacher in another subject. The method can yield TVA estimates which compare teachers within-school due to transitivity: comparisons can be made between teachers who are not directly connected through shared students, as long as they are linked through intermediary teachers. Exploiting teacher mobility across schools, it can also yield TVA estimates which compare teachers across schools. In schools with complete networks, I propose to measure the TVA of Math teachers using within-Literature teacher variation, relying on the Literature score of a student as control of the student's overall ability level—or the ability that is common between the two subjects. This network method yields unbiased TVA estimates as long as students are sorted to teachers based on their overall ability, rather than subject-specific differences in ability: in other words, how much better they are in Math than in Literature.

[1] See, for instance, Rothstein (2010), Hanushek (2011), Kane, McCaffrey, Miller, and Staiger (2013), Bacher-Hicks, Kane, and Staiger (2014), Jackson (2018), Kraft (2019), Bau and Das (2020).

[2] E.g. Hoxby and Leigh (2004), Jackson and Bruegmann (2009), Jackson (2013), Biasi (2021), among others.

Finally, for settings similar to the U.S. where lagged test scores are available, I also introduce a "hybrid" methodology, which incorporates lagged scores into the network estimator. The latter has an advantage over the network method as it does not assume a lack of sorting on subject-specific ability. It also has an advantage over standard methods, as it allows to control for time-varying student unobservables which may be correlated with teacher sorting.

I demonstrate that both the network and hybrid methods perform well in identifying the true teacher effect parameters in Monte Carlo simulations, in terms of the standard deviation of the effect, the mean squared error, the rank correlations, and the statistical relationship between the estimates and the true parameters. For the network method, this is the case if the identification assumption is satisfied—under no sorting on the difference between the lagged Math and Literature test scores of students. In addition, I use plausible simulations of teacher-student sorting driven by changes in student unobservables over time—such as students who become more motivated are assigned to better teachers—to show that the network and hybrid estimators are unbiased in such scenarios, while the TVA estimator commonly used in the literature (Kane & Staiger, 2008) is biased.

I then apply the method to observational data, using exhaustive administrative data from both New York City and France.[3] The use of New York City data has several important advantages for validating the network and hybrid methods. Because of the fact that students in New York City sit exams after every middle school year, I can apply all three methods—the network, hybrid and standard methods—to the data. This allows to assess the performance of the network and hybrid methods compared to the standard estimators in the field. I show that the estimators perform similarly to the standard TVA estimator (Kane & Staiger, 2008), based on the estimated standard deviation of the teacher effect, the rank correlations, squared differences, and statistical relationships with the standard TVA estimates. In addition, the availability of lagged test scores allows me to confirm the validity of the identification assumption of the network estimator in realized data. Specifically, I show that while teacher-student sorting on the average lagged test scores in Math and Reading is somewhat common (occurring systematically in about 10% of schools), sorting of Math teachers on the difference between Math and Reading scores, which proxies for a student's Math-specific ability, is very rare (occurring systematically in only 2% of schools).

Equipped with this validation of the network method, I provide novel evidence of the TVA distribution in France—an example of a setting where lagged test scores are not available—using French middle school teachers in Math and French. In France, more than 90% of teachers belong to a complete school network, making the estimates representative of the average teacher effect. Specifically, I show that a 1 standard deviation increase in TVA within a school improves student scores by 0.10 SD in Math and 0.07 SD in French. These

[3]Note that I use the subject "Literature" interchangeably with "Reading" and "French", in order to put the two subjects under the same general umbrella.

implied teacher effects are lower than the ones identified for the U.S. (Chetty et al., 2014a). These results highlight substantial variation in teacher effects across countries and suggest a potential discussion on the teacher practices that may be driving these differences.

This paper contributes to the extensive literature on parametric TVA estimation in three key ways.[4] First, I extend existing TVA methodologies (Rockoff, 2004, Kane & Staiger, 2008, Chetty et al., 2014a) by proposing a new method that does not rely on and can therefore generalize unbiased estimation of TVA to other countries than the U.S. I demonstrate that the network estimator can generate unbiased TVA estimates that closely match the true teacher parameters using cross-sectional test score data, rather than requiring panel data. This significantly broadens the applicability of TVA research to many countries with reliable administrative data, particularly in Europe and South America, where panel test score data is often unavailable. While cross-sectional student test scores across subjects have been widely used in other contexts—such as studying the effects of instructional time (Lavy, 2015) or linking teacher credentials and assessment scores to student achievement (Clotfelter, Ladd, & Vigdor, 2010; Benhenda, 2018)—this the first paper that leverages them for TVA estimation.

Second, the paper contributes to the literature by developing a hybrid estimator that extends the standard TVA framework in settings with annual standardized testing, such as the U.S. This approach not only addresses the conventional concerns of nonrandom teacher-student sorting—by controlling for lagged scores—but also captures previously overlooked forms of sorting driven by contemporaneous changes in student characteristics. By leveraging both time-invariant and time-varying student unobservables, this estimator can yield unbiased TVA estimates in settings in which standard methods would fail due to more complex sorting on time-varying student characteristics.

Third, this paper provides the first large-scale TVA estimates for France, offering novel insights into teacher effectiveness in a system where standardized testing is infrequent. By demonstrating the feasibility of TVA estimation in such contexts, this study can facilitate research on teacher quality in different countries, and on the pedagogical and institutional drivers of teaching quality (for instance, Rockoff, 2004, Rivkin, Hanushek, & Kain, 2005, Staiger, Gordon, & Kane, 2006, Kane, Rockoff, & Staiger, 2008, Rockoff, Jacob, Kane, & Staiger, 2011, Harris & Sass, 2014, Wiswall, 2013, JP & MA, 2015, Bold et al., 2016, Lavy, 2016, Bietenbeck, Piopiunik, & Wiederhold, 2018, Bau & Das, 2020). The evidence highlights potential cross-country variation in teacher effects and calls for a discussion on the teaching practices that may explain these differences. These findings contribute to the broader debate on how teacher effectiveness varies across educational systems and inform policies on teacher evaluation and professional development beyond U.S.-centric frameworks.

[4]More recently, there has also been growing interest in non-parametric TVA estimation (Gilraine, Gu, & McMillan, 2023).

The paper proceeds as follows. Section 2 discusses the statistical model and assumptions necessary to derive the network estimator and shows how to estimate it in the data. Section 3 compares the performance of the network, hybrid and traditional methods to the true teacher quality parameters in simulated data. Section 4 gives an overview of the New York City and French data used for the empirical test of the method. Section 5 focuses on the empirical results, starting from comprehensive tests of non-random sorting and finishing with the variation of TVA measured in both settings. Finally, Section 6 concludes.

2 Statistical model

In this section, I propose a simple framework of student test scores and teacher value-added. For simplicity of exposition, I assume there is only one period. I discuss the modifications to the framework when multiple time periods are introduced in Appendix A.1.

2.1 Notation and setup

Students Let the set of students be denoted $\mathcal{N}$, with a representative student i. Each student i is allocated to a school s from the set of all schools $\mathcal{S}$. Student i attends courses on Math and Literature, $z \in \{M, L\}$, during the school year and passes standardized exams in both subjects at the end of the school year. Student i is endowed with certain idiosyncratic qualities for passing the exams - one can think of intrinsic ability, motivation, interest, health, parental involvement, and many others. For simplicity, I refer to them broadly as "abilities" and distinguish between two types of abilities: one that is subject-specific, λ_i^z, and one that is common across subjects γ_i. Finally, i is endowed with a set of observable characteristics which affect i's performance in the exams, $\mathbf{X}_i^z$, which could be either subject-specific or common across subjects, such as family characteristics, absences from school, advanced courses taken, characteristics of class peers and school peers. The student's performance on the exam also depends on the class they have been assigned to within the school and the quality of their teachers.

Classes Let $\mathcal{C}$ represent the set of all classes. Let student i be sorted to a class $c(i, z)$ for each subject z by the school principal. For the purpose of notation simplicity, assume that $c(i, M) = c(i, L) \equiv c(i)$: the set of students sorted together in Math are also sorted together in subject Literature. A less restrictive alternative is to define $c(i, M, L)$ as the subgroup of students that are sorted together in Math and Literature. This definition would be useful for educational settings in which students are sorted into different peer groups in different subjects.

Each class experiences idiosyncratic class-specific shocks in each subject, $\zeta_{c(i)}^z$, that affect the performance of students in the respective subject. A simple example of such shock is

that students belonging to class $c(i)$ had to take the exam in a noisy environment due to roadworks close by.

Teachers Each class $c(i,t)$ is assigned a teacher for subject M and a teacher for subject L. Let $\mathcal{J}_z$ be the set of teachers in subject z, such that $j(z) \in \mathcal{J}_z$ denote a single teacher j in subject z. It follows that the set of all teachers is $\mathcal{J} \equiv \mathcal{J}_M \cup \mathcal{J}_L$. For notational simplicity, I denote a representative Math teacher $j(M) \equiv m$ and a representative Literature teacher $j(L) \equiv l$, respectively. For two subjects $M \neq L$, classroom $c(i)$'s teachers $m(c) \neq l(c)$. This setting therefore resembles a middle or high school allocation of teachers to classrooms, rather than an elementary school allocation.

Each teacher has a specific productivity, or value-added, denoted as $\mu_{j(z)}$, which is constant across classes and years. In Appendix A.1.1 I show the implications for TVA in a setting with more than one time period, with either an idiosyncratic year-specific shock to TVA or an experience-based year-specific shock to TVA.

Student exam scores I consider a standard model of student achievement with education inputs that have additively separable effects. Let student i's test score in Math, A_i^{*M}, be given by:

$$A_i^{*M} = b^z \times \mathbf{X}_i^M + \mu_m + \nu_i^M \tag{1}$$

such that b^M is a vector that represents the impact of observable characteristics $\mathbf{X}_i^M$ on student i's score, μ_m is the effect of i's Math teacher m, and ν_{im}^M is an error term. The latter can be decomposed into:

$$\nu_i^M = \lambda_i^{*M} + \gamma_i^* + \zeta_{c(i)}^M$$

where λ_i^{*M} and γ_i^* represent the ability components (Math-specific and common across subjects, respectively) measured with some error, and $\zeta_{c(i)}^M$ represents the class-specific shock for i's class.[5] The expression for the decomposition of i's Literature test score is analogous.

We are interested in the unobserved μ_m. The problem is that μ_m may be correlated with the other unobserved components of student i's score A_{im}^{*M}. The largest source of concern in the literature has been the potential correlation with the unobserved ability of the student, the sum of λ_i^M and γ_i. If teachers are allocated to students based on student ability, and one does not control for student ability, estimates of TVA would be biased. Traditional methods for estimating TVA use lagged test scores in standardized exams from the previous academic year to proxy for ability (e.g. Rockoff, 2004, Kane & Staiger, 2008, Chetty et al., 2014a, among others).

[5] This is to reflect the fact that student i's performance may not reflect their ability fully due to specific factors on the day of the exam, such as having fever.

In the absence of previous test scores, one can use the contemporaneous scores from the standardized exam in Literature to control for the ability component γ_i which is common across subjects. Such a within-student across-subject method is equivalent to taking the first difference of equation 1 across subjects M and L. Denoting $\Delta A_i^{*ML} \equiv A_i^{*M} - A_i^{*L}$:

$$\Delta A_i^{*ML} = (b^M \times \mathbf{X}_i^M - b^L \times \mathbf{X}_i^L) + (\mu_m - \mu_l) + (\nu_i^M - \nu_i^L) \tag{2}$$

where

$$\nu_i^M - \nu_i^L = (\lambda_i^{*M} - \lambda_i^{*L}) + (\zeta_{c(i)}^M - \zeta_{c(i)}^L)$$

The challenge for obtaining unbiased estimates using equation 2 is that the difference in error terms $(\nu_i^M - \nu_i^L)$ may be correlated with μ_m and μ_l.

ASSUMPTION 1 (Zero conditional mean of the error term) *The difference in error terms $(\nu_i^M - \nu_i^L)$ has a mean of zero conditional on the teacher effects $\{\mu_m\}_{m \in J_M}$ and $\{\mu_l\}_{l \in J_L}$ and the observable student characteristics $\mathbf{X}_i^z$. This constitutes:*

(a) $\mathbb{E}\left[\lambda_i^z | \mu_m, \mu_l, \mathbf{X}_i^M, \mathbf{X}_i^L\right] = 0$, $\forall i$, $\forall z\{M, L\}$, $\forall m \in \{J_M\}$, and $\forall l \in \{J_L\}$

(b) $\mathbb{E}\left[\zeta_{c(i)}^z | \mu_m, \mu_l, \mathbf{X}_i^M, \mathbf{X}_i^L\right] = 0$, $\forall c(i)$, $\forall z\{M, L\}$, $\forall m \in \mathcal{J}_M$, and $\forall l \in \mathcal{J}_L$

In other words, what Assumption 1(a) requires is that there is no difference in subject-specific ability, conditional on the baseline TVA of the two teachers who teach student i and observables. More specifically, it must not be the case that teacher m or teacher l is sorted to students who are on average better (or worse) in Math than in Literature. Testing and accounting for such sorting is key for obtaining unbiased estimates.

Assumption 1(b) requires that there is no difference in the class-level noise, conditional on the baseline TVA of the two teachers who teach class $c(i)$ and observables. Using the example of the noisy environment in which a class has to take the exam, the assumption would be violated if classrooms that draw negative shocks (e.g. are in a noisy environment due to construction) are sorted to the less (or more) experienced teachers, which is an unlikely event.[6]

An additional assumption is necessary on the behavior of the difference in the error terms (similarly to most prior studies on TVA).

ASSUMPTION 2 (Random sampling) *The difference in error terms $(\nu_i^M - \nu_i^L)$ is i.i.d. conditional on the teacher effects $\{\mu_m\}_{m \in J_M}$ and $\{\mu_l\}_{l \in J_L}$ and the observable student characteristics $\mathbf{X}_i^z$.*

[6]Note that the statistical model on the determinants of student test scores follows the literature, in that it does not assume any spillover effects of teachers—Math teachers do not influence Literature scores and vice versa. Appendix A.5 discusses the implications of potential spillover effects for the identification of TVA.

Finally, under satisfied Assumptions 1 and 2, the remaining issue with estimating equation 2 is dimensionality. As every student i in class c, $\forall i$, $\forall c \in \{C\}$, shares the same Math and Literature teachers as their classmates, the teacher effects are identified only through variation in test scores *across* classes. Consequently, if each teacher is assigned to only one class—resulting in C number of Math teachers and C number of Literature teachers—the total number of teacher effects to estimate, $2 \times C$, would exceed the number of classes, leading to a lack of sufficient variation to estimate the teacher effects reliably.

I show that the dimensionality problem is solved with the use of connected sets of teachers, which I call *networks* of teachers.

DEFINITION (Teacher networks): *For a subset of classes $\mathcal{C}^n \subseteq \mathcal{C}$, a network of teachers is a subset of teachers $\mathcal{J}^n \subseteq \mathcal{J}$, for which for each class $c \in \mathcal{C}^n$, there exists at least one class $c' \in \mathcal{C}^n \setminus \{c\}$ such that the teacher for subject z in class c also teaches in class c':*

$$\exists c' \in \mathcal{C} \setminus \{c\},\ j(c, z) = j(c', z).$$

In other words, a network of teachers is a group of teachers who are connected through the classes they teach. This can be a direct connection, where the same teacher teaches both classes c and c', or an indirect connection—through transitivity. From Assumption 1 it follows that all teachers in a network have on average students with the same subject-specific abilities.

To illustrate the network identification, Figure 1 proposes two example of connected sets within a school. The blue nodes are different Math teachers, whereas the yellow nodes are Literature teachers. If an edge connects a Math and a Literature teacher, the Math and Literature teachers teach in the same class at least once.

The set of teachers in subfigure (a) all belong to the same network, as every teacher can be connected to all other teachers through their classroom observations. For instance, Math teachers M_1 and M_2 are both observed with the Literature teacher L_1. Similarly, this is the case for Math teachers M_1 and M_6, who are observed with Literature teacher L_4. Teachers M_2 and M_6 are not observed with the same Literature teacher, but as they are both connected to teacher M_1 through their classroom observations with L_1 and L_4, respectively, they are also connected to each other through transitivity.

Note that in this example, the available classroom observations exactly match the number of total teachers, such that all teacher effects can be estimated. Appendix A.2 includes a formal proof that the dimensionality problem is solved within a connected set, which allows for the OLS estimator to be unique and unbiased at finite distance.

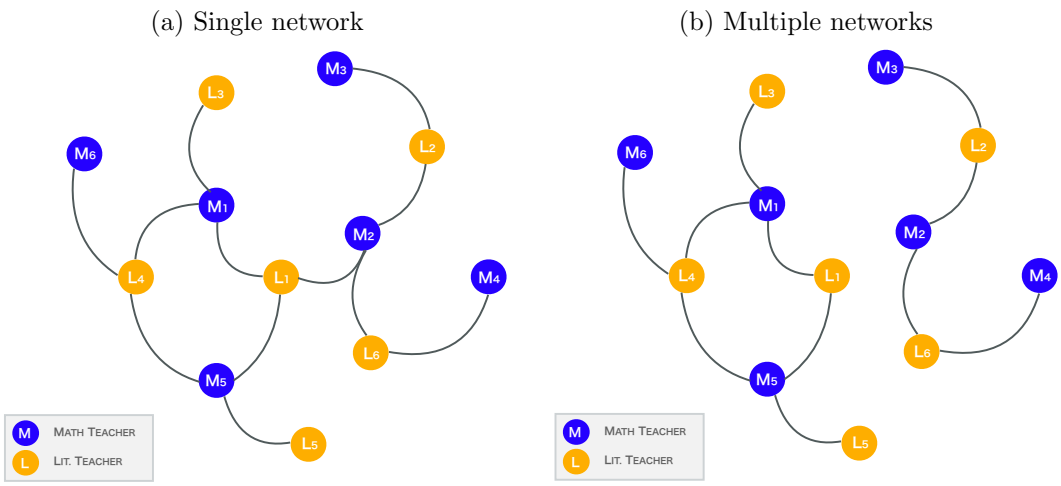


Figure 1: Example networks of teachers within a school

Note: The figures represent two examples of networks present within a representative school, for a set of Math and Literature teachers. The sequence of teachers is not specific to a single year. Each node represents a teacher - either in Math (blue) or in Literature (yellow). Each edge between two teachers signifies that the two teachers share at least one class.

Alternatively, subfigure (b) shows an example of a school with two distinct networks of teachers. In such a school, there are more teachers than classes available to estimate all teacher effects within the school. It follows that while teacher effects can be estimated as relative measures of TVA within each of the two networks (following the proof in Appendix A.2), comparisons between the two networks within the schools is not possible without an additional assumption on the relative TVA across the two connected networks.

However, such an assumption may be implausible if there is reason to believe that there is sorting of teachers to classes across networks. For example, if connected sets span smaller than the set of teachers within a school, certain networks might include more experienced teachers, while others might have novice teachers, such that the former might have a higher likelihood of being placed in front of higher-ability students, or vice versa. Therefore, if the sorting of teachers to classes is non-random, we expect that there are systematic differences between teachers in different networks. When comparing teachers across connected sets, these systematic differences can bias the results, as differences between sets could be driven by factors unrelated to teacher performance, such as school policies or student demographics.

Certain networks are more natural to compare teachers within than others. Two types of networks are particularly interesting: the school-level networks and the largest networks.

A school-level network for school s—a connected set that spans the universe of teachers in school s (similarly to the represented school in Figure 1 (a))—ensures that the network of teachers in a given school can be compared on the same scale. Within a school-level connected set, the teachers are subject to the same institutional conditions, meaning that differences

in observed performance are more likely to reflect true differences in teacher value-added.

It follows that the effects of teachers in different schools are not comparable to each other, as they are only a relative measure of teacher effects *within each school*, for every school with a school-level connected set. A within-school comparison of TVA is preferable in cases where students and teachers are sorted non-randomly across schools, in ways that cannot be controlled for with observable school characteristics.

Alternatively, the largest network - the connected set that includes the largest number of teachers from $\mathcal{J}$ - uses the mobility of teachers across schools in order to expand the network of teachers. The largest possible connected set is the universe of teachers itself, $\mathcal{J}$. It follows that, in this case, within the large network, TVA would be an absolute measure of teacher quality, as all teachers in the universe of schools can be put on the same distribution of TVA. However, the plausibility of Assumption 1 would be small if students and teachers were not randomly sorted across schools. For instance, teachers in wealthier schools may systematically perform better due to factors unrelated to their individual teaching abilities, such as better resources or higher-achieving students. Some such factors can be controlled for, such as lagged student test scores, but if there are reasons to believe that not all differences in schools can be explained by observables, across-school measures of TVA should be used with caution.

2.2 Network estimator

Using the outlined identification strategy, we can estimate TVA for the set of teachers belonging to a network, such that the networks of interest can either be at the level of the school for a within-school TVA estimation (in which case teachers are given unique school-specific identifiers) or at the level of the largest network for an across-school TVA estimation. Equation 2 can be estimated by OLS as:

$$\Delta A_i^{*ML} = \alpha + \mathbf{X}_i^M \beta^M + \mathbf{X}_i^L \beta^L + \sum_{m \in \mathcal{J}_M^n} \mu_m \mathbb{1}(j(M) = m) + \sum_{l \in \mathcal{J}_L^n} \mu_l \mathbb{1}(j(L) = l) + \varepsilon_i \quad (3)$$

where α is an intercept which captures the fact that students might be better in Math than in Literature ceteris paribus, ΔA_i^{*ML} is the difference between student i's Math and Literature scores, $\mathbf{X}_i^M$ and $\mathbf{X}_i^L$ are observable characteristics, which could include both student-level and class-level characteristics, μ_m are Math teacher fixed effects, μ_l are Literature teacher fixed effects, and ε_i is the idiosyncratic error term.

The estimation strategy bases the identification of TVA as Math teacher effects are identified within the same Literature teacher, and vice versa for Literature fixed effects. Appendix A.3 provides the intuition behind the equivalence.

Teacher fixed effects are added to the regression rather than left in the error term as teacher quality may be correlated with observable student characteristics. If this is the case and

one does not identify the βs within Math and Literature teachers, estimates of the βs would overstate or understate the impact of the $\mathbf{X}$s because a part of teacher quality is attributed to them (Chetty et al., 2014a).

To compute the TVA estimates for Math teachers, I take the residual from the regression 3 that purges the effect of the Literature teacher effects[7] and observed covariates from ΔA_i^{*ML}, and denote it $\widehat{A}_i^{ML}$, according to equation 2:

$$\widehat{A}_i^{ML} \equiv \widehat{\mu}_m + \widehat{\varepsilon}_i$$

As teachers are often observed to teach for only a few years and tend to teach few classrooms per year, a common problem with school data is that while unbiased, the estimators $\widehat{\mu}_m$ is likely noisy, and especially so for teachers with less experience. To deal with this issue, I shrink TVA estimates using an Empirical Bayes shrinkage method adapted from the literature. The Empirical Bayes estimator weights each teacher observation by its precision to obtain a precision-weighted average of teacher effects, and then weights each teacher's teacher effect by an estimate of the effect's reliability (the signal-to-noise ratio).

Specifically, using the teacher-year-average residuals for each m teacher $\widehat{A}_{mt}$, I form a weighted-average residual $\widehat{A}_m$ per teacher m, which is the minimum variance unbiased estimate of μ_m for each teacher:

$$\widehat{A}_m = \sum_t w_{mt}\widehat{A}_{mt} \text{ where } w_{mt} = \frac{h_{mt}}{\sum_t h_{mt}} \text{ and } h_{mt} = \frac{1}{Var(\widehat{A}_{mt}|\mu_m)}$$

Finally, using $\widehat{A}_m$ I construct an empirical Bayes estimator for each teacher's TVA by multiplying this weighted average residual by an estimate of its reliability (the signal-to-noise ratio):

$$\widehat{TVA}_m = \widehat{A}_m \left(\frac{\sigma_\mu^2}{\sigma_\mu^2 + 1/\sum_t h_{mt}} \right)$$

where σ_μ^2 is the variance of the Math teacher effect.

To estimate σ_μ^2, I first compute the covariance between the teacher's yearly residual in two randomly selected years, weighted by the number of students in each year, similarly to Kane and Staiger (2008) and Chetty et al. (2014a):

$$Cov(\widehat{A}_{mt}, \widehat{A}_{mt'})$$

Then, to account for any remaining variance of the link teacher in the residual, I also compute the covariance between the link teacher's yearly residual in the same way:

$$\sigma_l^2 = Cov(\widehat{A}_{lt}, \widehat{A}_{lt'})$$

[7] Note that we do not need to include the estimated Literature teacher effects in this residual, as by definition they are orthogonal to the Math teacher effects.

Finally, I compute the variance of the fixed teacher component as:

$$\sigma_\mu^2 = Cov(\widehat{A}_{mt}, \widehat{A}_{mt'}) - \sigma_l^2$$

In Appendix A.4, I outline the full Empirical Bayes shrinkage method used and show how I compute the variances for each component of the residual in a setting which does not assume zero year-specific teacher noise.

Note that as the resulting estimates are shrunk towards zero, the Bayesian Shrinkage estimator is no longer an unbiased predictor of the TVA of teacher m. However, it is forecast unbiased - it is the best predictor of future performance of m, as it has higher accuracy due to reduced mean squared error (for proof, see Appendix A.4.1).

2.3 Hybrid method

The network method's advantage of controlling for unobserved factors at time t may be beneficial even in settings where controlling for lagged test scores is possible, as is the case in the United States. In particular, combining the network and traditional methods into a "hybrid" method would allow to control for time-varying unobserved student characteristics which may be important for the way teachers are sorted to students.

To see this, let us now assume that student ability varies over time, and we can proxy for the unobserved ability of student i at time t, $a_{it}^z = \lambda_{it}^z + \gamma_{it}$ using past test scores, $A_{i,t-1}^{*z}$ This is possible under the implicit assumptions that, (1) ability is an autoregressive process of the type:

$$a_{it}^z = \phi a_{it-1}^z + \epsilon_{it}^z$$

with ϵ_{it}^z being a random noise, and (2) test scores of i in year $t-1$, $A_{i,t-1}^{*z}$, are a good proxy for a_{it-1}^z and therefore for a_{it}^z.

If this is the case, Assumptions 1 and 2 derived in the previous sections cease to be necessary. We can estimate the TVA of Math teachers following the existing literature under standard assumptions for unbiasedness, starting from the OLS regression:

$$A_{it}^* = \alpha + \mathbf{X}_{it}^M \beta^M + \sum_{m \in \mathcal{J}_M^n} \mu_m \mathbb{1}(j(M) = m) + \varepsilon_{it} \tag{4}$$

where $\mathbf{X}_{it}^M$ contains $A_{i,t-1}^{*M}$ (and could also include $A_{i,t-1}^{*L}$). Similarly to before, we can use the residual:

$$\widehat{A}_{it} = \widehat{\mu}_m + \widehat{\varepsilon}_{it}$$

to derive the Empirical Bayes estimator of TVA as before. Under the statistical model specified in equation 1, this estimator would be forecast unbiased.

Let us now modify one of these implicit assumptions. Specifically, assume instead that

$$a^z_{it} = \phi a^z_{it-1} + \Delta\pi_{it}$$

such that π_{it} is no longer random noise, but is instead a change in circumstances of i between periods t and $t-1$. For the sake of example, let us think of π_{it} as the level of motivation of i at time t.

Let us now imagine that students are (re-)sorted to classrooms and teachers yearly based their ability a^z_{it}, which now takes into account the level of motivation they promise at the start of each school year t, π_{it}. If this is the case, the estimator obtained using equation 4 will now be biased, because it fails to control for π_{it}.

Instead, a *hybrid method* - a method that uses the network identification but also controls for lagged scores, would be able to control for the sorting and obtain forecast unbiased estimates. This is because by taking the difference in contemporaneous scores $A^{*M}_{it} - A^{*L}_{it}$, π_{it} would cancel out, similarly to the common ability component.

Specifically, we can modify equation 3 to include lagged scores:

$$\begin{aligned}\Delta A^{*ML}_{it} =& \alpha + A^M_{i,t-1}\gamma^M + A^L_{i,t-1}\gamma^L + \mathbf{X}^M_{it}\beta^M + \mathbf{X}^L_{it}\beta^L \\ &+ \sum_{m\in\mathcal{J}^n_M} \mu_m \mathbb{1}(j(M)=m) + \sum_{l\in\mathcal{J}^n_L} \mu_l \mathbb{1}(j(L)=l) + \varepsilon_{it}\end{aligned}$$

The same Empirical Bayes shrinkage can be applied in this case.

Advantages compared to the network method The hybrid method has an obvious advantage over the network method, because being able to control for lagged scores ensures that Assumption 1 is no longer needed for unbiasedness, as we can plausibly control for subject-specific ability.

Advantages compared to the literature The hybrid method's advantage over traditional methods is that it allows to control for student i × time t variation.

Note that a more complex process of a^z_{it} which takes into account subject-specific motivation π^z_{it} would still lead to bias in case of sorting of students to teachers based on π^z_{it}, as the latter would require to control for information at the student i × time t × subject z level.

3 Comparative performance in simulated data

I use Monte Carlo simulations to compare the finite-sample performance of the Kane and Staiger (2008) (hereafter KS) method estimator, the network method estimator and the hybrid method estimator relative to a benchmark distribution of TVA (infeasible in practice), which is the true distribution of teacher quality. I focus on estimating within-school TVA

distributions of Math teachers for 1,000 simulated schools with the average school characteristics in my sample of schools (details on this setup in Appendix B.1).

The data-generating process follows the statistical model of student scores described in Section 2: student test scores are a linear function of student ability (common and subject-specific), teacher quality, classroom noise, and student noise.[8] The choice of parameters is guided by the findings of Chetty et al. (2014a) and additional necessary assumptions regarding the shares of common ability, subject-specific ability, and student noise in the student variance, as well as the persistence of abilities over time (detailed breakdown provided in Appendix B.2).

I consider the performance of these estimators based on (i) scatterplots between estimates and true TVA, (ii) rank correlations of estimates and true TVA, (iii) mean squared error (MSE) of the estimates and true TVA, and (iv) the estimated SD of TVA.

To produce the scatter plots with the true TVA parameters, I use the pre-shrinkage estimates $\widehat{A}_m$ of each method, as by definition the shrunk estimates introduce bias, to determine how closely they fit the 45-degree line. I also include the coefficients obtained from a regression between the two estimates. These graphs illustrate by construction the unbiasedness of the estimates under the satisfied identification assumptions.

To produce the rank correlations, as more unreliable estimates would be shrunk more, thus changing the ranking of teachers within a school, I use the pre-shrinkage estimates to determine the rank of a teacher within their school. I then compute school-specific rank correlations of the estimates and the true TVA parameters.

To compare the MSE produced by each method, I take instead the shrunk estimates $\widehat{TVA}_m$ which are forecast unbiased and thus have the highest accuracy for predicting individual-level TVA. I compute MSEs for each teacher and take the average MSE per school, comparing the distribution of MSEs across methods.

To compare the SD of TVA predicted by each method, I take the estimated covariance σ^2_μ as the predictor of the true variation in TVA (as done in Kane & Staiger, 2008 and Chetty et al., 2014a).

I do these tests under four scenarios of sorting between teachers and students. First, random sorting of students to classrooms and random assignment of teachers to classrooms. Second, sorting of students to classrooms based on their lagged average score in Math and Literature, and positive assignment of Math teachers to classrooms. Third, sorting of students to classrooms based on their lagged Math scores, and positive assignment of Math and Literature teachers to classrooms. Fourth, sorting of students to classrooms based on the change in

[8] For simplicity, I assume observable characteristics (other than lagged scores) are already residualized from student scores.

their unobserved factors (e.g. motivation, ability, or studiousness), and positive assignment of Math teachers to classrooms. The four cases of sorting are chosen as illustrative of the power and weaknesses of the network and hybrid estimators compared to the KS estimator.

The case of random sorting Figure 3 plots the relationship between each estimate $\widehat{A}_m$ and the true TVA parameters μ_m within school for the case of random sorting, dividing the TVA parameters into 100 equal-sized groups and plotting the mean value of $\widehat{A}_m$ in each bin. The regression coefficient and standard error reported are estimated on the micro-data rather than the binned averages, controlling for school fixed effects, with standard errors clustered at the school level. The fit is almost perfect for all three estimators, such that the regression coefficient is not statistically different than 1.

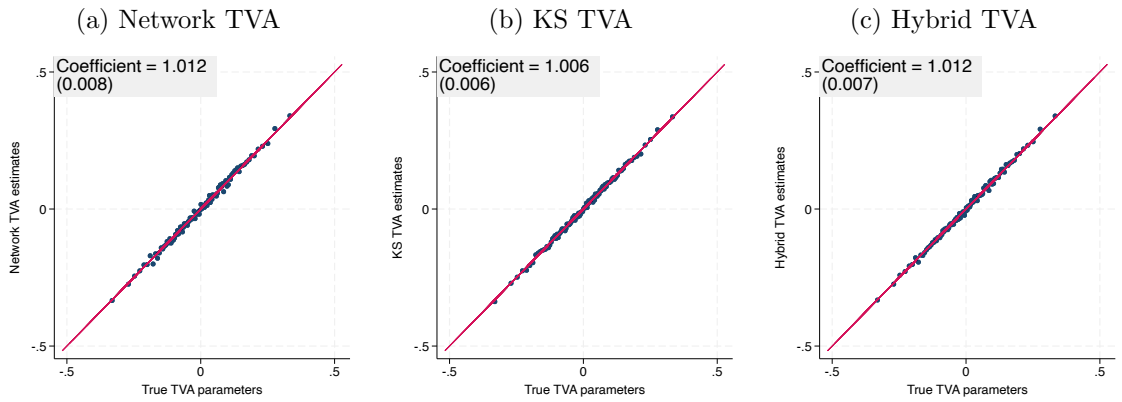


Figure 3: Relationship between estimates of TVA and true TVA parameters

Note: The figures are based on the analysis where sorting of teachers to students is random. Each binscatter represents the relationship between the TVA estimates (on the y-axis) and the true TVA parameters (on the x-axis) within school, such that Panel (a) reports the results for the Network estimator, Panel (b) reports the results for the KS estimator, and Panel (c) reports the results for the Hybrid estimator. The TVA parameters are divided into 100 equal-sized groups and each dot represents the mean value of $\widehat{A}_m$ in each bin. The red line is the 45-degree line. The gray boxes in each panel reports the regression coefficient and standard error from a regression of $\widehat{A}_m$ on μ_m, controlling for school fixed effects, rather than the binned averages. Standard errors are clustered at the school level.

All three distributions of school-specific Spearman correlation with the true TVA parameters are statistically similar according to the confidence intervals reporting the 10th and 90th percentile in the distributions (Figure 5 Panel (a)). The KS estimator has a slight advantage, with the highest median correlation (0.88), followed by the hybrid (0.86) and the network (0.81). Similarly, all three distributions of school-specific MSEs are insignificantly different from each other (Figure 5 Panel (b)). The KS estimator has the smallest median MSE (0.0025), followed closely by the hybrid (0.003) and the network estimators (0.0039). Finally, all three estimators predict the SD of the true TVA distribution very closely (Figure 5 Panel (c)). In particular, for a 0.1331 true SD of TVA, the KS has the closest predicted variation

of 0.1353, followed by the network (0.1298) and hybrid (0.1291) methods. The empirical distributions of the estimators are slightly less dispersed than the true TVA distribution.

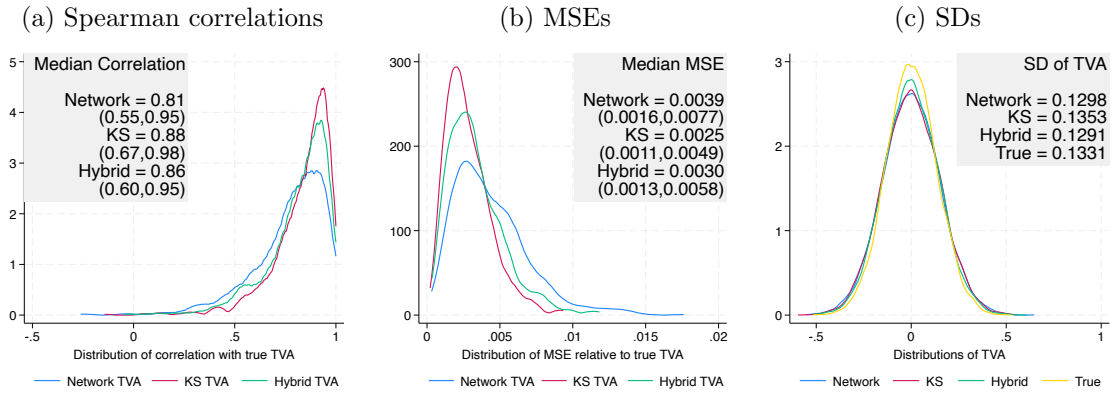


Figure 5: Predictive power of estimates compared to true TVA

Note: The figures are based on the analysis where sorting of teachers to students is random. Panel (a) depicts the distribution of Spearman correlations between the Network, Hybrid or KS estimates of TVA and the true TVA parameters respectively, computed at the school level. The gray box shows the correlation for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (b) depicts the distribution of mean squared errors (MSEs) of the Network, Hybrid or KS estimates of TVA compared to the true TVA parameters respectively, averaged at the school level. The gray box shows the MSE for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (c) shows the empirical distributions of the TVA estimators and the distribution of true TVA. The gray box shows the covariance of yearly teacher residuals for each method and the SD of true TVA.

The case of sorting on lagged average score In the case of positive sorting on lagged average scores in Math and Literature, the estimation error increases slightly for all estimators, leading to slightly higher regression coefficients and standard errors for the relationship with true TVA parameters (Figures B.2). Nevertheless, the relationship is insignificantly different than 1. The distributions of school-level Spearman correlations and MSEs are slightly more dispersed but not significantly different than the ones in scenario 1 (see Panels (a) and (b) in Figure B.4). Finally, the distributions of TVA are almost identical to before, with the KS method's predicted variation increasing slightly, while the hybrid method's predicted variation decreasing slightly. Overall, there are no significant changes in conclusions compared to the scenario of random sorting.

The case of sorting on the difference between lagged Math and Literature scores As a reminder, in the case of positive assignment of teachers to students based on the difference between lagged Math and Literature scores of students, the identification assumption of sorting is broken and the network estimator is biased upwards. On the other hand, the hybrid and KS estimators still converge to the true TVA parameters. The bias of the

network estimator is shown in Figures B.6: the regression coefficient is now significantly higher than 1 and, as seen in the scatterplot in Panel (a) as well, the network estimates are on average larger compared to the true TVA for estimates above 0 and smaller for estimates below 0. The larger error is also shown in the distribution of MSEs (Figure B.8 Panel (b)) and the increase in the estimated variation of TVA (Figure B.8 Panel (c)), which for the network estimator now increases to 0.3409. The trade-off of the introduced positive bias is that the more dispersed distribution of TVA makes it easier to rank teachers correctly, thus increasing the median Spearman correlation above both the hybrid and KS (0.95).

The case of sorting on changes in ability As a reminder, in cases of sorting on the change in unobserved student factors between $t-1$ and t, the KS estimator is biased. On the other hand, the network and hybrid estimators still converge to the true TVA parameters. It follows that there are virtually no changes to the binscatters (Figure B.6) and the distributions of Spearman correlations, MSEs and TVA for the network and hybrid methods (see Figure B.12). For the KS method, the regression coefficient is now significantly higher than 1 and, as seen in the scatterplot in Panel (b) as well, the network estimates are on average larger compared to the true TVA for estimates above 0 and smaller for estimates below 0.

For the KS estimator, the median MSE increases significantly, which drives the estimated SD of TVA to increase significantly, too (0.264). The introduced positive bias increases the median Spearman correlation significantly, similarly to the previous case (0.98).

Discussion on alternative cases Note that the identification assumption is not broken if students are sorted non-randomly to their classrooms but teachers are randomly assigned to classrooms. It is also not broken if the link (Literature) teachers are assigned non-randomly to students but Math teachers are randomly assigned, given the assumption of no spillover effects.

In cases where both Math and Literature teachers are assigned non-randomly, conclusions about bias would follow directly the conclusions outlined above, where only the Math teacher is assigned non-randomly. However, the more extreme this type of sorting is, the smaller the probability of observing complete networks, as Math and Literature teachers are observed with fewer and less diverse counterparts.

Finally, negative assignment - such that better teachers are assigned to lower-performing students - is another possibility. Note that the conclusions about bias would follow directly those outlined above, but the sign of the bias would be negative. In cases that lead to biased estimators, negative sorting would lead to a smaller estimated variation of TVA than the true, making it harder to predict TVA correctly, therefore leading to lower Spearman correlations as well. If, alternatively, sorting on the lagged Math scores is positive for Math teachers but negative for Literature teachers, the bias in the network estimator is lower.

Summary of findings Under a satisfied identification assumption, the network and hybrid methods perform relatively similarly to the KS method, with moderately larger median MSEs and moderately smaller Spearman correlation with the true parameters, but not statistically different. The differences in MSE and Spearman correlation are small enough to make the network estimator useful in settings without annual standardized testing.

Having said that, if one believes that sorting of students to teachers might be based on student factors that vary between $t-1$ and t (as in the example of the last case), the network method becomes better at identifying unbiased TVA. In such cases, the hybrid method outperforms both alternatives in terms of MSE and predicted SD of TVA, though the network method remains a close second.

It follows that the hybrid method is the only method that produces unbiased estimates in all 4 representative cases of sorting. Therefore, if one is unsure of the type of sorting that may occur within schools and has lagged test scores available (as is the case in the United States), the advantage of the hybrid method is important to distinguish.

4 Data

The data in this study draws from two exhaustive administrative datasets with matched and detailed student-teacher data - the New York City public school data and the French public school data. I focus exclusively on 8th-graders in New York City data, in order to mirror the available test score data on 9th-graders in France.

New York City The New York City dataset covers 8th-grade Math and English teachers and their students between the school years 2003-2004 and 2008-2009.[9] The dataset has information on Math and English scores in the 8th grade and all previous grades starting from elementary school, which I restandardize to mean-zero and a standard deviation of 1 per subject and year. Tests are administered statewide as required under the No Child Left Behind law. Furthermore, the dataset contains detailed information on ethnicity, gender, age, limited English proficiency, receipt of free lunch, receipt of special education services, participation in honors classes, absences from class, suspensions, foreign nationality, repetition of grade, and repetition of test.

I drop all students for which either Math or Literature scores (either current or past ones) are missing, all classrooms that have the same Math and Literature teacher or more than one teacher in either subject, and all schools that have only one Math/Literature teacher teaching the 8th grade throughout the sample period (as identification is based on within-school variation). Following Chetty et al. (2014a), I also exclude students who receive instruction

[9] I restrict the data post-2003 as identifiers for Math and English teachers stabilize to roughly 85% after 2003, according to Chetty et al. (2014a).

at home, as well as classrooms where more than 25% of students receive special education services, and classrooms with less than 10 or more than 50 students. I also exclude teachers linked to more than 200 students in a single grade, as Chetty et al. (2014a) conclude these are likely to be mislinked to classrooms or teachers.

For both Math and Literature, this leads to about 80% of the sample of schools having complete networks, or about 250 middle schools with a total of 170,000 8th graders[10] (with about 75% of teachers being part of a complete network, representing roughly 1000 teachers).[11]

France The French dataset covers 9-th grade Math and French teachers and their students between 2006-2007 and 2020-2021. I focus on Metropolitan France. The dataset has information on Math and French scores in the 9th grade, but has no previous score information per student, as annual standardized testing does not exist.[12] I restandardize all test scores ot mean-zero and a standard deviation of 1 per subject and year.

The tests are administered nationwide as part of the National Brevet Diploma (DNB) necessary to finish middle school. The DNB diploma takes into account both continuous assessment (worth 57%) and standardized assessment (worth 43%). These exams test the knowledge, skills and culture acquired throughout middle school in Math, French language and other subjects.

In addition to exam scores, the dataset contains detailed student information on gender, age, city of birth, receipt of needs-based scholarship, parental or guardian occupation (from which a socio-economic status category can be deduced) and labor market activity, nationality, repetition of grade, repetition of test, and option classes chosen.

As for New York City data, I drop all students for which either Math or French scores are missing, all classrooms that have the same Math and French teacher or more than one teacher in either subject, and all schools that have only one Math/French teacher teaching the 9th grade throughout the sample period.

In France, this leads to about 83% of the sample of schools having complete networks, or about 4900 middle schools[13] with a total of 5.2 million 9th graders (with about 90% of teachers in each subject being part of a complete network, representing roughly 25,000

[10] Note that the total number of 8th graders is 335,000. About 58,000 students are dropped as they have not taken one of the two exams. About 2,000 are dropped because their Math teacher is also their Literature teacher. About 60,000 are dropped because they are taught by a Math or Literature teacher with only 1 year of experience within a school.

[11] Further restricting the data to teachers with more years of experience within a school increases the share of teachers and schools within a complete network, at the expense of having fewer teachers with identifiable TVA, see Figure **??**.

[12] Starting from the year 2018-2019, standardized tests are also administered to 6th-graders in Math and French.

[13] Or high schools where the 9th grade is being taught.

teachers in Math and 27,000 teachers in French).

5 Results

5.1 Tests of sorting

Assumption 1 can be directly tested using observable characteristics, with tests similar to those proposed by Rothstein (2009). It is, however, first worth discussing the plausibility of Assumption 1 in different educational settings, and in particular two settings that may create concerns for the identification.

First, a concern in many settings would be that students may be sorted to schools non-randomly based on unobserved relative ability which we cannot plausibly control for. Let school s specialize in teaching Math, such that all Math teachers have high TVA and all students have high relative Math ability. If the analysis is conducted within school, such that we only compare teachers relatively within s, this would not be a concern. However, for across-school comparisons, it may be. As long as this specialization does not change between year t and year $t-1$, one can use the school-level average scores in Math and Literature from previous years to plausibly control for such non-random sorting. However, if the school only begins such type of sorting in year t (e.g. with the arrival of new high TVA Math teachers), one would tend to underestimate the TVA of the Math teachers in s.

Second, another concern with Assumption 1 may be related to the differential sorting of students to classrooms in different subjects at certain schools. If a student i is sorted to a classroom $c(i, M)$ in Math and $c(i, L)$ in Literature, the probability that the principal based this sorting on subject-specific ability is higher. This may be explicit - for instance, based on students signing up for Advanced Math rather than a regular Math class. If this is the case, the analysis can be conducted within-stream. It is less clear whether the method is applicable in cases where such type of sorting is not explicit, and depends on tests of the identifying assumption in the specific setting.

Lagged scores and current teachers I test for sorting on lagged scores in NYC schools using the fact that lagged scores cannot be explained by a student's current teacher for any other reason but sorting.

Importantly, the identifying assumption does not require having no sorting, it requires that students are not sorted more based on their Math test scores compared to their Literature test scores, conditional on the set of controls. Student controls are important because they may capture some of this sorting of students to teachers. Classroom controls are also important because what we would like to avoid capturing is just that students are sorted non-randomly into a classroom. More specifically, non-random student sorting to classrooms is a necessary

but not sufficient condition for bias; for bias to occur, non-random student sorting needs to be coupled with non-random assignment of teachers to classrooms (a "specialization" of teachers, such that better teachers are always found in better classrooms).

Students are very unlikely to repeat/skip a year in the hope of being allocated to a specific teacher. For this reason, students may only be non-randomly sorted to a classroom within their cohort. Teachers on the other hand may be sorted to teach either to different grade level in a given year based on the ability of students in specific grades, or to a different class within a grade level based on student ability, provided that they teach in the specific year. I test for both types of non-random sorting within school, focusing exclusively on schools with complete networks, as the rest of the schools are not in the analysis, and are more likely to have non-random sorting.

I test the validity of the sorting assumption in two steps. I first regress lagged student scores on current observable characteristics of students and their classrooms:

$$A^{*z}_{i,t-1} = \mathbf{X}_{it}\beta + \theta_s + \varepsilon^z_{i,t-1}$$

where $\mathbf{X}^z_{it}$ includes student controls and classroom-level controls for the year t classroom of student i, and θ_s are school fixed effects.

I take the estimated residual of $\varepsilon^z_{i,t-1}$ and denote it $A^z_{i,t-1}$, and proceed by estimating separately for each school:

$$A^z_{i,t-1} = \mu_{j(z,t)} + \varepsilon^z_{i,t-1} \tag{5}$$

where $\mu_{j(z,t)}$ are the set of teacher coefficients such that each teacher has a per-year per-school coefficient.

The relevant statistics we can examine is the p-value on an F-test for joint significance of the teacher coefficients (as in Rothstein, 2010) within the year in the particular school.

If schools are large enough, one would expect a perfectly uniform distribution of p-values. As pointed out by Rothstein (2010), one may be worried that in grades with very few classrooms per year, these tests might be misleading, as even if assignment is in reality random, there may still be substantial overlap in the composition of classrooms between year t and $t-1$. This would lead to a larger mass of p-values that indicate significant non-random assignment. For this reason, I conduct additional Monte Carlo simulations where I re-assign teachers to classrooms randomly in two ways. First, I take only 8th-grade classrooms and re-assign 8th-grade teachers to random classrooms within the year, preserving the number of classrooms each teacher teaches in the year (Type 1 sorting). Second, I take classrooms from all grades and, once again preserving the number of classrooms each teacher teaches in the year, I reassign teachers to classrooms. I then take only 8th-grade classrooms for the analysis

(Type 2 sorting). I report in the appendix the difference between the true distribution of school-level p-values from the F-test against the placebo distribution.

In the case of NYC, in both types of sorting, we can see that the true and placebo distributions almost perfectly overlap (Figure C.14). This is also visible from the fact that in both types of sorting, only an excess of about 2% of school-year combinations seems to show signs of non-random sorting in the true data, compared to the simulated data (Figure C). By contrast, there is an excess of between 9% and 10% of school-years that show signs of non-random allocation on the average score in the true data, compared to the simulated data. This is consistent with the idea that sorting on the average Math-Literature score is much more likely.

Network vs hybrid estimates Another test of non-random sorting which can be conducted in the case of New York City due to the availability of annual standardized testing is the comparison of the network and hybrid estimates. As both types of estimations are identical with the sole exception of the addition or omission of lagged test scores, comparing the obtained estimates from the two methods would allow us to say whether the omission of lagged test scores causes bias in the network estimates. The coefficient of a regression of the network estimates on the hybrid estimates is statistically indistinguishable from 1 (as shown in Figure 7 Panel (c)). This crude test further confirms the validity of the main identifying assumption.

5.2 Estimates of TVA

I estimate TVA in both the sample of teachers in public schools based in NYC and in France, using the methodology outlined in Section 2.2.

The vectors of student controls $\mathbf{X}_{it}^{z}$ and $\mathbf{X}_{it}^{z'}$ include student-level, classroom-level and, in the case of estimating absolute TVA, school-level controls.

Specifically for France, student-level controls that vary by student i include the socio-economic status (SES) of student i, their age, gender, and means-based scholarship and advanced classes taken. Controls that vary by subject and student include exam repetitions. Classroom-level controls include all student-level controls averaged at the classroom-level, e.g. number of peers, average SES, average age, percentage of females. As I focus on estimating TVA within school, I do not add any school-level controls.

For NYC, student-level controls that vary by student i include the ethnicity of student i, their age, gender, a dummy for benefitting from a free lunch, a dummy for being in special education, a dummy for having limited English knowledge, a dummy for having repeated a grade, number of absences and suspensions in the year, and a dummy for being a foreign student. Controls that vary by subject include whether or not they have repeated the exam,

and a dummy for being in an honours program. As for the analysis for France, classroom-level controls include all student-level controls averaged at the classroom-level.

Standard deviation of TVA I compute the SD of TVA as the covariance between yearly observations of a teacher m (l, respectively), corrected for the covariance of yearly observations of the link teacher l (m, respectively).

Specifically, in the context of New York City, I find that the SD of TVA in Math is 0.1356 according to the network estimator and 0.1311 according to the hybrid estimator. For Literature, the SD of TVA is 0.1025 according to the network and 0.0978 according to the hybrid method. It follows that both methods produce very comparable results, indeed confirming that any positive bias due to non-random sorting on lagged scores in the subject in question is limited.

For France, where I can only obtain the network estimator, a 1 SD increase in Math (French) TVA leads to a 0.098 SD (0.065 SD) increase in Math (French) scores. These results indicate that on average, the effect of teachers on standardized test scores is lower in France than it is in NYC.

Table 1: SD of TVA by method and subject

Note: This table reports the estimated standard deviation (SD) of TVA using two different estimators—Network and Hybrid—in both NYC and the French setting, for Math and Literature. The SD is computed as the square root of σ^2_μ, where $\sigma^2_\mu = Cov(\widehat{A}_{mt}, \widehat{A}_{mt'}) - \sigma^2_l$, such that $\sigma^2_l = Cov(\widehat{A}_{lt}, \widehat{A}_{lt'})$.

	Math		Literature	
	Network	Hybrid	Network	Hybrid
NYC	0.1356	0.1311	0.1025	0.0978
France	0.098		0.065	

Comparison to KS estimates using New York City data It is worth comparing the point estimates obtained by the network, hybrid and KS methods in the setting of NYC, where lagged test score data is available. Contrary to comparisons in simulated data, where comparisons of each method's estimates are made with regards to the true TVA parameters, comparisons in real data can only be done across estimation methods. As each method's estimates contain a certain amount of error, one cannot expect that the network and hybrid estimates are identical to the KS method's estimates. However, as they are identifying the same teacher effects, they should lay within the same band.

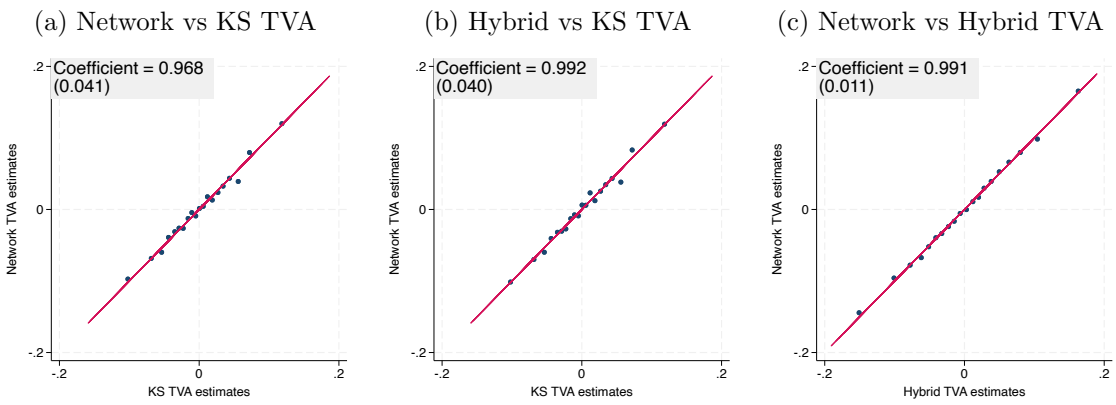


Figure 7: Relationship between estimates of TVA

Note: Each binscatter represents the relationship between two different TVA estimates on the x- and y-axes within school, such that Panel (a) reports the results for the Network vs KS estimator, Panel (b) reports the results for the Hybrid vs KS estimator, and Panel (c) reports the results for the Network vs Hybrid estimator. The TVA parameters are divided into 20 equal-sized groups and each dot represents the mean value of $\widehat{TVA}_m$ in each bin. The red line is the 45-degree line. The gray boxes in each panel report the regression coefficient and standard error from a regression of $\widehat{TVA}_m^E$ on $\widehat{TVA}_m^{E'}$ (where E and E' are two different estimators), controlling for school fixed effects, rather than the binned averages. Standard errors are clustered at the school level.

Figure 7 depicts the binscatters of the KS estimates against the (a) network and (b) hybrid estimates, respectively, and (c) the binscatter of the network against the hybrid estimates. For panel (a) and (b), the binscatters are constructed by dividing the KS estimate $\widehat{A}_{j(z)}$ into 20 equal-sized groups (vingtiles) and plotting the means of the (a) network and (b) hybrid estimates, respectively, within each bin against the mean value of the KS estimates $\widehat{A}_{j(z)}$ within each bin. Similarly, for panel (c), I divide the hybrid estimates into 20 vingtiles and plot the mean of the network estimates within each bin. Each of the methods' estimates are highly comparable, depicted by the fact they are clustered around the 45-degree line. The coefficients quoted come from regressions of the network and hybrid method's estimates on the KS estimates, controlling for school fixed effects. For the network (hybrid) estimates, the slope of 0.968 (0.992) is not significantly different from 1 at the 95% confidence level, signifying that the estimates produced by each method are on average equal. The network and hybrid estimates are also very comparable (with a slope of 0.991), reflecting the fact that there is little sorting of teachers to students based on the lagged Math scores.

While the binscatters are useful for the relationship between estimates on average, it is worthwhile to compare the specific point estimates. I do this in two ways, similar to the tests performed in simulated data.

First, I compute the school-level Spearman correlation between the (i) network and (ii) hybrid estimates, and the KS estimates. Figure 9 panel (a) plots the distribution of school-level correlations, while the numbers reported in the top left corner report the median correlation

for each of the two methods, with the correlations of the schools at 10th and 90th percentile in parentheses. Both distributions look very similar, with the largest mass of the correlations falling above 60%: the median correlation for the network method stands at 73%, while that of the hybrid method - at 80%. Given the estimation noise, these correlations are not significantly different than those found in the simulated data when comparing estimates to true TVA parameters.

Second, I compute the school-level MSE between the (i) network and (ii) hybrid estimates, and the KS estimates. As the KS estimates are only estimates of the true TVA, these measures of MSE can be interpreted as mean squared differences between the network and hybrid method, and the KS method. Figure 9 panel (b) plots the distribution of school-level MSEs, while the numbers reported in the top right corner report the median MSE for each of the two methods, with the MSE of the schools at 10th and 90th percentile in parentheses. Once again, the two distributions look very similar, with the hybrid MSE distribution having a slightly smaller MSE both at the median and in the tails. Specifically, the median MSE for the network method stands at 0.0017, while that of the hybrid method - at 0.0014.

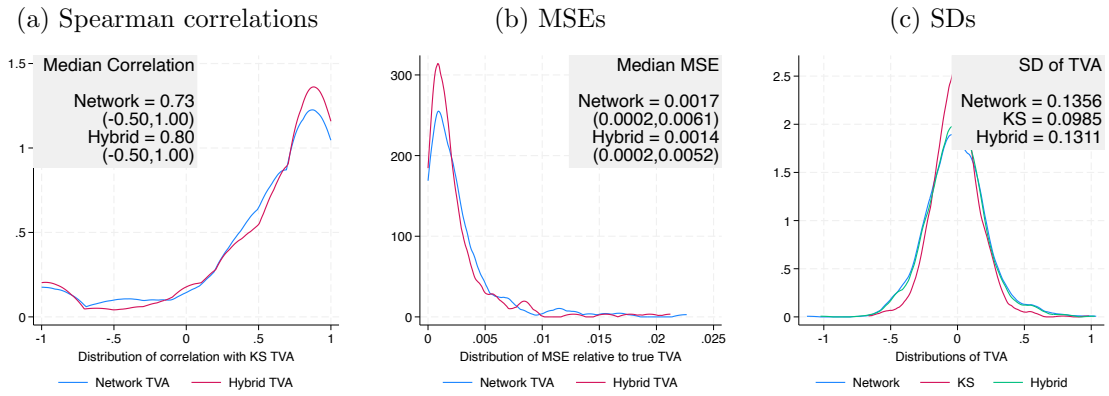


Figure 9: Predictive power of Network and Hybrid estimates compared to KS estimates of TVA

Note: Panel (a) depicts the distribution of Spearman correlations between the Network or Hybrid estimates of TVA and the KS estimates of TVA respectively, computed at the school level. The gray box shows the correlation for the median school for each of the two estimates with the KS estimates respectively, and the interval of the 10th and 90th percentile. Panel (b) depicts the distribution of mean squared errors (MSEs) of the Network or Hybrid estimates of TVA compared to the KS estimates of TVA respectively, averaged at the school level. The gray box shows the MSE for the median school for each of the two estimate with the KS estimates respectively, and the interval of the 10th and 90th percentile. Panel (c) shows the empirical distributions of the TVA estimators. The gray box shows the covariance of yearly teacher residuals for each method.

Finally, I compare the empirical distributions of TVA produced by each method and the estimated effect of teachers. Consistent with the findings of Chetty et al. (2014a), I show that the Kane and Staiger (2008) estimator produces a SD of TVA which is biased towards

zero in true data. Nevertheless, the methods produce roughly comparable results.

Overall, these comparisons do confirm that the network method can be a good substitute to standard methods in scenarios where lagged scores are not available.

6 Conclusion

This paper introduces a novel methodology for estimating teacher value-added (TVA) in settings where students do not take standardized tests annually, addressing a fundamental limitation of existing TVA models. Traditional approaches rely on lagged test scores to control for student ability, which is not viable in most education systems outside the U.S., where standardized testing is less frequent. My method circumvents this challenge by leveraging within-student, across-subject variation and networks of teachers to estimate TVA in the absence of panel data of student test scores. This framework expands the applicability of TVA estimation beyond the U.S. context and provides a feasible alternative for evaluating teacher effectiveness in education systems where standard TVA models are infeasible.

The network-based methodology exploits the natural structure of teacher assignments across subjects to derive plausibly unbiased TVA estimates. By focusing on schools with complete teacher networks, I show that Math and Literature scores can serve as controls for students' overall ability, enabling TVA estimation without the need for prior test scores. I further introduce a hybrid approach that integrates elements of both network-based and traditional TVA models. This method refines the standard TVA estimation by accounting for time-varying student unobservables, mitigating biases introduced by sorting mechanisms that assign students to teachers based on evolving traits such as motivation or effort. Through Monte Carlo simulations, I demonstrate that both the network and hybrid methods perform well in identifying the true distribution of teacher effects, achieving accuracy levels comparable to traditional TVA estimators under satisfied identification assumptions.

Applying these methods to administrative data from New York City, where panel data is available, I validate their effectiveness in real-world settings. I compare my estimates to those obtained using standard TVA methods and show that the network and hybrid estimators yield similar distributions of teacher effects to that produced by the standard method. Additionally, I confirm the key identification assumption of the network estimator—namely, that if students are sorted to teachers non-randomly, they are sorted based on their overall ability rather than subject-specific differences—by showing that sorting based on Math-Literature score differences occurs in only a small fraction of schools.

In France, where traditional TVA estimation is infeasible due to the lack of annual testing, I provide the first evidence of TVA distributions among middle school teachers. My findings reveal that teacher effects in France are generally lower than those reported in the U.S.,

raising important questions about cross-country differences in teacher practices and their impact on student outcomes.

Beyond its methodological contributions, this paper has important implications for education policy. The feasibility of TVA estimation without lagged scores opens the door for broader applications in teacher evaluation, particularly in contexts where standard TVA methods cannot be implemented. Moreover, the hybrid approach offers a more robust framework for addressing student-teacher sorting mechanisms that are often overlooked in traditional TVA models. By demonstrating that these methods yield unbiased estimates under reasonable assumptions, this paper provides a foundation for policymakers and researchers seeking to measure teacher effectiveness in a wider range of educational settings.

References

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in los angeles* (Tech. Rep. No. 20657). NBER Working Paper.

Bau, N., & Das, J. (2020). The impact of individual teachers on student achievement: Evidence from panel data. *Americal Economic Journal: Economic Policy*, *12*(1), 62–96.

Benhenda, A. (2018). Teacher Screening, On the Job Evaluations and Performancee. *Quantitative Social Science - UCL Social Research Institute, University College London*(18-06).

Biasi, B. (2021). The labor market for teachers under different pay schemes. *American Economic Journal: Economic Policy*, *13*(2), 65–102.

Bietenbeck, J., Piopiunik, M., & Wiederhold, S. (2018). Africa's Skill Tragedy: Does Teach- ers' Lack of Knowledge Lead to Low Student Performance? *Journal of Human Resources*, *53*, 553–578.

Bold, T., Filmer, D. P., Martin, G., Molina, E., Christophe Rockmore, J. S., Brian W. Stacy, & Wane, W. (2016). What Do Teachers Know and Do? Does It Matter? Evidence from Primary Schools in Africa. *World Bank Working Paper 7956*.

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

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

Clotfelter, C. T., Ladd, H., & Vigdor, J. (2010). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *Journal of Human Resources*, *45*(3).

de Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, *110*(9), 2964–96.

Gilraine, M., Gu, J., & McMillan, R. (2023). A Nonparametric Approach for Studying Teacher Impacts. *NBER Working Paper 27094*.

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

Harris, D., & Sass, T. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, *40*, 183–204.

Hoxby, C., & Leigh, A. (2004). Pulled Away or Pushed out? Explaining the Decline of Teacher Aptitude in the United States. *The American Economic Review*, *94*(2), 236–240.

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics*, *95*(4), 1096–1116.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85–108.

JP, P., & MA, K. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119.

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

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631.

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

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

Lavy, V. (2015). Do Differences in Schools' Instruction Time Explain International Achievement Gaps in Math, Science, and Reading?: Evidence from Developed and Developing Countries. *The Economic Journal*, *125*(588), 397–424.

Lavy, V. (2016). What makes an effective teacher? Quasi-experimental evidence. *CESifo Economic Studies*, *62*(1), 88–125.

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 Papers and Proceedings*, *94*(2), 247—252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*, 43–74.

Rothstein, J. (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, *4*(4), 537–571.

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

Staiger, D. O., Gordon, R., & Kane, T. J. (2006). Identifying Effective Teachers Using Performance on the Job. *Brookings Report*.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, *100*,

61–78.

Appendix

A Statistical model details

A.1 Setting with multiple time periods

A.1.1 Adaptation of the model

To extend the simple setting introduced in Section 2, we can rewrite the statistical model in equation 1 as:

$$A_{it}^{*M} = b^z \times \mathbf{X}_{it}^M + \mu_m + \nu_{it}^M$$

where the error term ν_{it}^M can be decomposed into:

$$\nu_{it}^M = \lambda_{it}^{*M} + \gamma_{it}^* + \zeta_{c(i)}^M + \omega_{mt}$$

such that λ_{it}^{*M} and γ_{it}^* represent the year-specific ability components (Math-specific and common across subjects, respectively) measured with some error, $\zeta_{c(i)}^M$ represents the class-specific shock for i's class, and ω_{mt} represents the year-specific teacher shock.

There are two alternative assumptions we could take regarding the transitory component of TVA.

ASSUMPTION A.1. (Stationarity of transitory TVA) *The transitory component of TVA, ω_{mt} follows a stationary process:*

$$\mathbb{E}[\omega_{mt}] = 0, \text{ and}$$
$$\text{(i) } cov(\omega_{mt}, \omega_{mt+\tau}) = 0 \text{ and } var(\omega_{mt}) = \sigma_\omega \text{ for all } t \text{ and } \tau \neq 0.$$

Assumption A.1. allows us to treat the transitory component of TVA, ω_{mt} as random and uncorrelated. It follows that the only change in equation 3 is the additional components in the error term:

$$\Delta A_i^{*ML} = (b^M \times \mathbf{X}_i^M - b^L \times \mathbf{X}_i^L) + (\mu_m - \mu_l) + (\nu_i^M - \nu_i^L)$$

where

$$\nu_i^M - \nu_i^L = (\lambda_i^{*M} - \lambda_i^{*L}) + (\zeta_{c(i)}^M - \zeta_{c(i)}^L) + (\omega_{mt} - \omega_{lt})$$

Alternatively, one can instead assume the transitory component is a linear function of the teacher's years of experience at year t. Let the number of years of experience of m at t be denoted $E_{m(t)}$.

ASSUMPTION A.2. (Experience-based transitory TVA) *The transitory component of TVA, ω_{mt}, is a linear function of m's years of experience at year t and random teacher-specific noise that follows a stationary process:*

$$\omega_{jm} = \mathbb{1}(E_{m(t)} = E) + \delta_{mt}$$

where

$$cov(\delta_{mt}, \delta_{mt+\tau}) = 0 \text{ and } var(\delta_{mt}) = \sigma_\delta \text{ for all } t \text{ and } \tau \neq 0.$$

Equation 3 needs to be modified to thus include both the years of experience of both teachers and the additional components in the error term:

$$\Delta A_i^{*ML} = (b^M \times \mathbf{X}_i^M - b^L \times \mathbf{X}_i^L) + (\mu_m - \mu_l) + (E_{mt} - E_{lt}) + (\nu_i^M - \nu_i^L)$$

where

$$\nu_i^M - \nu_i^L = (\lambda_i^{*M} - \lambda_i^{*L}) + (\zeta_{c(i)}^M - \zeta_{c(i)}^L) + (\delta_{mt} - \delta_{lt})$$

It follows that Assumptions 1 and 2 in Section 2 would have to be adapted to account for the additional observable component of experience:

ASSUMPTION 1* (Zero conditional mean of the error term) *The difference in error terms $(\nu_i^M - \nu_i^L)$ has a mean of zero conditional on the teacher effects $\{\mu_m\}_{m \in J_M}$ and $\{\mu_l\}_{l \in J_L}$, the observable student characteristics $\mathbf{X}_i^z$ and the years of experience of the two teachers $E_{m(t)}$ and $E_{l(t)}$. This constitutes:*

(a) $\mathbb{E}\left[\lambda_{it}^z | \mu_m, \mu_l, \mathbf{X}_{it}^M, \mathbf{X}_{it}^L, E_{m(t)}, E_{m(t)}\right] = 0$, $\forall i$, $\forall z\{M, L\}$, $\forall m \in \{J_M\}$, $\forall l \in \{J_L\}$, and $\forall t$

(b) $\mathbb{E}\left[\zeta_{c(i)}^z | \mu_m, \mu_l, \mathbf{X}_i^M, \mathbf{X}_i^L, E_{m(t)}\right] = 0$, $\forall c(i)$, $\forall z\{M, L\}$, $\forall m \in \mathcal{J}_M$, $\forall l \in \mathcal{J}_L$, and $\forall t$

ASSUMPTION 2* (Random sampling) *The difference in error terms $(\nu_i^M - \nu_i^L)$ is i.i.d. conditional on the teacher effects $\{\mu_m\}_{m \in J_M}$ and $\{\mu_l\}_{l \in J_L}$, the observable student characteristics $\mathbf{X}_i^z$ and the years of experience of the two teachers $E_{m(t)}$ and $E_{l(t)}$.*

Consequently, the estimation of TVA by OLS would be modified to:

$$\begin{aligned}\Delta A_i^{*ML} =& \alpha + \mathbf{X}_i^M \beta^M + \mathbf{X}_i^L \beta^L + \sum_{m \in \mathcal{J}_M^n} \mu_m \mathbb{1}(j(M) = m) + \sum_{l \in \mathcal{J}_L^n} \mu_l \mathbb{1}(j(L) = l) \\ &+ \sum_{y \in Y(m)} \alpha_{ym} \mathbb{1}(E_{m(t)} = E_y) + \sum_{y \in Y(l)} \alpha_{yl} \mathbb{1}(E_{l(t)} = E_y) + \varepsilon_i\end{aligned}$$

where $Y(z)$ are the total number of distinct years of experience observed in the group of z teachers. Alternatively, $Y(z)$ can be the total number of experience bands in the group of z teachers.

A.1.2 Yearly networks of teachers

One might also be interested in comparing networks of teachers exclusively within the same year. This might be useful if there are significant changes to exams over the years. Figure A.1 shows why this may be difficult to achieve especially in smaller schools.

Imagine a school that teaches students in 5 different years. If we want to compare TVA within that school only within year, we can "split" Figure 1 into 5 subfigures, such that each subfigure now shows only the teachers who teach 9th-grade math and 9th-grade literature in the specific year.

In Year 1, the two math teachers observed, M_1 and M_2, are in a network. The same is true in Year 5, with math teachers M_1, M_2 and M_5. However, the sets of teachers in Year 2, 3 and 4 and not connected. It follows that for the example school, we may only uncover relative TVA for math teachers within Year 1 and Year 5. The same logic follows for literature teachers. It therefore makes sense to not restrict the within-school analysis to within-year, in order to increase the coverage of school-level connected sets.

Finally, notice that within-year analysis is not possible if we compare teachers across schools, as teachers are in most cases not mobile within a year. Specifically, mobile teachers are more likely to move from school A, where they teach until year $t-1$, to school B in year t. Therefore, such variation may only exist if teachers teach in more than one school in the same school year or they move mid-year, both of which cases are much less common.

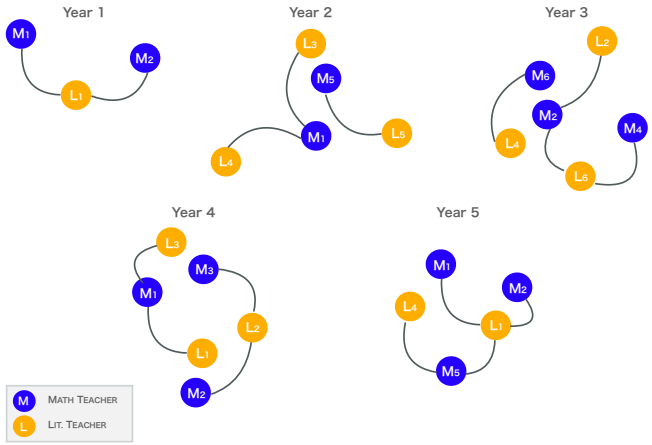


Figure A.1: Yearly school allocations and connected sets

Note: This figure represents the sets of teachers within a random school per year. The sets are connected only in Year 1 and Year 5, thus showing why the method is more widely applicable when school-level networks are allowed to exist both within and across years.

A.2 Networks and dimensionality

In this section, I provide a proof that the dimensionality problem outlined in Section 2 can be solved with the use of connected sets, called networks, through proving the absence of perfect multicollinearity between regressors within networks.

Let us consider a network which extends to the level of a school s, such that all teachers can be linked to each other through their classroom observations with teachers from another subject.

For simplicity of exposition and without a loss of generality, let teachers $m(1)$ and $l(1)$ who teach class $c(1)$ be the reference Math and Literature teachers, such that $\mu_{m(1)} = \mu_{l(1)} = 0$. This can be achieved by adding an intercept $\mathbb{1}$ to equation 2. It follows that $\mu_{m(1)}$ and $\mu_{l(1)}$ are excluded from the model in equation 2. Note that the same result would follow if no intercept are added such that the within-school mean of μ_m and μ_l, $\overline{\mu}_s^M$ and $\overline{\mu}_s^L$ are zero.

Consider a set of scalars $(\{a_m\}_{m\in[2,\overline{M}]}, \{b_l\}_{l\in[2,\overline{L}]}, d)$ such that:

$$\delta\mathbb{1} + \sum_{m=2}^{\overline{M}} a_m\mu_m + \sum_{l=2}^{\overline{L}} a_l\mu_l = 0 \tag{A.1}$$

It can be shown that the equality holds for each school s and all classes $c \in C$. Specifically, consider first the reference class $c(1)$. For reference class $c(1)$, $\mu_{m(1)} = \mu_{l(1)} = 0$ by normalization, which implies that $\delta = 0$.

By the definition of a network, it must be that if school s has a complete network, either $m(1)$ or $l(1)$ is observed with at least one other class. Without loss of generality, let this class be $c(2)$ such that $m(1)$ is observed in $c(2)$ with a Literature teacher $l(2)$. As $\mu_{m(1)} = 0$ and $\delta = 0$, from equation A.1 it follows that $a_{l(2)} = 0$. The logic can be extended to all other classes within the school s. For example, if $l(2)$ teaches in class $c(3)$ with another Math teacher $m(3)$, it would follow that $a_{m(3)} = 0$.

Therefore, $d = 0$ and $a_{m(c)} = b_{l(c)} = 0\ \forall c \in [2, C]$. This proves that $\mathbb{1}$, $\{\mu_m\}_{m\in[2,\overline{M}]}$, $\{\mu_l\}_{l\in[2,\overline{L}]}$ are not collinear. The OLS estimator is unique and unbiased at finite distance.

Consider instead that school s has two separate networks, such that there exists a set of classes C', which is not connected to the set of classes C. The set of classes C' is taught by Math teachers $[\overline{M}+1, \overline{M}+\overline{M}']$ and Literature teachers $[\overline{L}+1, \overline{L}+\overline{L}']$. Without further normalization, equation A.1 can be rewritten as:

$$\delta\mathbb{1} + \sum_{m=2}^{\overline{M}} a_m\mu_m + \sum_{l=2}^{\overline{L}} a_l\mu_l + \sum_{m=\overline{M}+1}^{\overline{M}'} b_m\mu_m + \sum_{l=\overline{L}+1}^{\overline{L}'} b_l\mu_l = 0 \tag{A.2}$$

Under $d = 0$, $\mu_{m(1)} = \mu_{l(1)} = 0$ and $a_{m(c)} = b_{l(c)} = 0\ \forall c \in [2, C]$, absence of perfect

multicollinearity would hold iff:

$$\sum_{m=\overline{M}+1}^{\overline{M}'} b_m \mu_m + \sum_{l=\overline{L}+1}^{\overline{L}'} b_l \mu_l = 0 \tag{A.3}$$

which is feasible only if $b_m = b_l = 0$ or under additional normalization, for instance $\mu_{m(\overline{M}+1)} = \mu_{l(\overline{L}+1)} = 0$. The latter would entail that identification is feasible within each of the two sets.

It follows that under assumptions 1-2 and with the use of networks, teacher effects are identified within network, i.e. the OLS estimation of $\{\mu_m\}_{m\in\{J_M\}}$ and $\{\mu_l\}_{l\in\{J_L\}}$, are uniquely defined and unbiased estimators of the teacher effects of finite distance:

$$\begin{aligned}
&\mathbb{E}[\widehat{\varepsilon}_{it}|\mu_m, \mu_l, \mathbf{X}_{it}] = \\
&\mathbb{E}\left[(\lambda_{it}^M - \lambda_{it}^L) + (\zeta_{c(i)}^M - \zeta_{c(i)}^L) + (\omega_m - \omega_l)|\mu_m, \mathbb{1}_l\mu_l, \mathbf{X}_{it}\right] = \\
&\mathbb{E}\left[(\lambda_{it}^M - \lambda_{it}^L)|\mu_m, \mu_l, \mathbf{X}_{it}\right] + \mathbb{E}\left[(\zeta_{c(i)}^M - \zeta_{c(i)}^L)|\mu_m, \mu_l, \mathbf{X}_{it}\right] + \\
&\mathbb{E}\left[(\omega_m - \omega_l)|\mu_m, \mu_l, \mathbf{X}_{it}\right] = 0
\end{aligned}$$

It follows that:

$$E(\widehat{\mu}_m|\mu_l, \mathbf{X}_{it}) = \mu_m, \forall m \in J_M \text{ and } E(\widehat{\mu}|\mu_m, \mathbf{X}_{itl}) = \mu_l, \forall l \in J_L$$

A.3 Equivalence between TWFE equation and FD for two Math and one Literature teachers

Note that for the case when $\mathcal{J}_M = 2$ and $\mathcal{J}_L \leq 2$, the equation 3 is numerically equivalent to a first-difference equation. To see this, take the example of for the case when $\mathcal{J}_M = \{m(1), m(2)\}$ and $\mathcal{J}_L = \{l\}$. Let student i in class $c(1)$ be taught by $m(1)$ in year t and student k in class $c(2)$ be taught by $m(2)$ in year t'. It follows that for i:

$$\Delta A_{it}^{*ML} = \alpha + \mathbf{X}_{it}^M \beta^M + \mathbf{X}_{it}^L \beta^L + \mu_{m(1)} + \mu_l + \varepsilon_{it}$$

and for k:

$$A_{kt'}^{*ML} = \alpha + \mathbf{X}_{kt'}^M \beta^M + \mathbf{X}_{kt'}^L \beta^L + \mu_{m(2)} + \mu_l + \varepsilon_{kt'}$$

Taking the first difference of these two differences and noting the fact that teacher l is teaching both i and k:

$$\Delta A_{it}^{*ML} - A_{kt'}^{*ML} = \left(\mathbf{X}_{it}^M - \mathbf{X}_{kt'}^M\right)\beta^M + \left(\mathbf{X}_{it}^L - \mathbf{X}_{kt'}^L\right)\beta^L + (\mu_{m(1)} - \mu_{m(2)}) + (\varepsilon_{it} - \varepsilon_{kt'})$$

It follows that we can isolate the relative effect of teacher $m(1)$ and teacher $m(2)$ by rearranging:

$$(\mu_{m(1)} - \mu_{m(2)}) = (\Delta A_{it}^{*ML} - A_{kt'}^{*ML}) - \left(\mathbf{X}_{it}^M - \mathbf{X}_{kt'}^M\right)\beta^M + \left(\mathbf{X}_{it}^L - \mathbf{X}_{kt'}^L\right)\beta^L - (\varepsilon_{it} - \varepsilon_{kt'}) \tag{A.4}$$

Equation A.4, averaged across all students i and k, would allow to uncover relative Math teacher effects of $m(1)$ and $m(2)$.

In the case of multiple Math and Literature teachers, equation 3 is no longer numerically equivalent to equation A.4. It is, however, equivalent to the weighted average of the full set of difference-in-differences equations (de Chaisemartin & D'Haultfœuille, 2020). It is important to note that, given the nature of class observations, we do not expect that every Math teacher is observed with each Literature teacher from the network, and vice versa (similarly to the illustrative example of a network). It thus follows that the available set of difference-in-differences equations which go in the estimation of the teacher fixed effects is smaller than the full set of potential difference-in-differences equations. As certain pairwise comparisons of teachers in the same subject are missing, some teacher fixed effects are identified based on the transitivity property, as long as teacher in question is part of the network. Importantly, all fixed effects are identified within the network the respective teachers belong to.

A.4 Empirical Bayes shrinkage

As teachers are often observed to teach for only a few years and tend to teach few classrooms per year, a common problem with school data is that while unbiased, the estimators $\widehat{\mu}_m$ is likely noisy, and especially so for teachers with less experience. To deal with this issue, I shrink TVA estimates using an Empirical Bayes shrinkage method adapted from the literature.

After estimating equation 3 by OLS, to compute the TVA estimates for Math teachers, I take the residual that purges the effect of the Literature teacher effects[14] and observed covariates from ΔA_i^{*ML}, and denote it $\widehat{A}_i^{ML}$, according to equation 2:

$$\widehat{A}_{it}^{ML} \equiv \widehat{\mu}_m + \widehat{\varepsilon}_{it}$$

We can use these sum of residuals to form empirical Bayes estimates for teachers' value-added. I do this in three steps.

First, I estimate the variances $\widehat{A}_{it}$ and each of its underlying parameters:

$$\widehat{A}_{it} = \overline{TVA}_m - \overline{TVA}_l + (\lambda_{it}^M - \lambda_{it}^L) + (\zeta_{c(i)}^M - \zeta_{c(i)}^L) + (\omega_{mt} - \omega_{lt})$$

Notably, I denote the variance of $\overline{TVA}_m$ as σ_μ^2, the (residual) variance of $\overline{TVA}_l$ as σ_l^2, the variance of $(\lambda_{it}^M - \lambda_{it}^L)$ as σ_λ^2, the variance of $(\zeta_{c(i)}^M - \zeta_{c(i)}^L)$ as σ_ζ^2, and the variance of $(\omega_{mt} - \omega_{lt})$ as σ_ω^2. Finally, I denote the total variance as σ_A^2.

I define the total variance as:

$$\sigma_A^2 = Var(\widehat{A}_{it}) \times \left(\frac{N-1}{N-K}\right)$$

[14] Note that we do not need to include the estimated Literature teacher effects in this residual, as by definition they are orthogonal to the Math teacher effects.

where N is the total number of students and K is the number of control variables in equation 3. The scaling term adjusts for the degrees of freedom (Chetty et al., 2014a).

I estimate the individual-level variance σ^2_λ as:

$$\sigma^2_\lambda = Var(\widehat{A}_{it} - \widehat{A}_{c(i)}) \times \left(\frac{N-1}{N-K-C+1}\right)$$

where $\widehat{A}_{c(i)})$ is the classroom $c(i)$-level average value of $\widehat{A}_{it}$, such that $(\widehat{A}_{it} - \widehat{A}_{c(i)})$ are within-classroom individual deviations from the mean residual and C is the total number of classes used in equation 3. The scaling term adjusts for the degrees of freedom (Chetty et al., 2014a).

To estimate the variance of the fixed teacher component, σ^2_μ, I first compute the covariance between the teacher's yearly residual in two randomly selected years, weighted by the number of students in each year, similarly to Kane and Staiger (2008) and Chetty et al. (2014a):

$$Cov(\widehat{A}_{mt}, \widehat{A}_{mt'})$$

Then, to account for any remaining variance of the link teacher in the residual, I also compute the covariance between the link teacher's yearly residual in the same way:

$$\sigma^2_l = Cov(\widehat{A}_{lt}, \widehat{A}_{lt'})$$

Finally, I compute the variance of the fixed teacher component as:

$$\sigma^2_\mu = Cov(\widehat{A}_{mt}, \widehat{A}_{mt'}) - \sigma^2_l$$

It follows that the estimated sum of the variance of the classroom component and the difference in transitory teacher effect between the Math and Literature teachers is the remainder:

$$var\left((\zeta^z_{c(i)} - \zeta^{z'}_{c(i)}) + (\omega_{mt} - \omega_{lt})\right) = \sigma^2_\zeta + \sigma^2_\omega = \sigma^2_A - \sigma^2_\mu - \sigma^2_\lambda$$

I then use the teacher-year-average residuals for each Math teacher $\widehat{A}_{mt}$ and form a weighted-average residual $\widehat{A}_m$ per teacher, which is the minimum variance unbiased estimate of μ_m for each teacher. Specifically, each teacher-year observation is weighted by its precision w_{mt} (the inverse of its conditional variance), such that years in which a teacher teaches more students receive more weight because they have a smaller variance:

$$\widehat{A}_m = \sum_t w_{mt}\widehat{A}_{mt} \text{ where } w_{mt} = \frac{h_{mt}}{\sum_t h_{mt}} \text{ and}$$
$$h_{mt} = \frac{1}{Var(\widehat{A}_{mt}|\mu_m)} = \frac{1}{\sigma^2_\zeta + \sigma^2_\omega + (\sigma^2_\lambda / N_{mt})}$$

Finally, using $\widehat{A}_m$ I construct an empirical Bayes estimator for each teacher's TVA by multiplying this weighted average residual by an estimate of its reliability (the signal-to-noise ratio):

$$\widehat{TVA}_m = \widehat{A}_m \left(\frac{\sigma^2_\mu}{\sigma^2_\mu + 1/\sum_t h_{mt}}\right) \tag{A.5}$$

A.4.1 Proof of forecast unbiasedness of Empirical shrinkage estimator

To see that the empirical Bayes network estimator is forecast unbiased, making use of Assumptions 1-2 and denoting for simplicity:

$$\left(\frac{\sigma_{\mu}^{2}}{\sigma_{\mu}^{2}+1/\sum_{t}h_{mt}}\right)\equiv k_{mt}$$

We can expand equation A.5:

$$\widehat{TVA}_{m}=\widehat{A}_{m}k_{mt}=\left(\sum_{t}w_{mt}\widehat{A}_{mt}\right)k_{mt}=\left(\sum_{t}w_{mt}\frac{\sum_{i\in C_{t}^{m}}N_{c(i)}\cdot\widehat{A}_{it}}{\sum_{i\in C_{t}^{m}}N_{c(i)}}\right)k_{mt}$$

Using the fact that:

$$\widehat{A}_{it}=\widehat{\mu}_{m}+\widehat{\varepsilon}_{it}$$

and that:

$$\frac{\sum_{i\in C_{t}^{m}}N_{c(i)}\cdot\widehat{\mu}_{m}}{\sum_{i\in C_{t}^{m}}N_{c(i)}}=\widehat{\mu}_{m}$$

we can simplify the equation for $\widehat{TVA}_{m}$ to:

$$\widehat{TVA}_{m}=\left(\sum_{t}w_{mt}\frac{\sum_{i\in C_{t}^{m}}N_{c(i)}\cdot(\widehat{\mu}_{m}+\widehat{\varepsilon}_{it})}{\sum_{i\in C_{t}^{m}}N_{c(i)}}\right)k_{mt}=$$
$$\left(\sum_{t}w_{mt}k_{mt}\widehat{\mu}_{m}\right)+\left(\sum_{t}w_{mt}\frac{\sum_{i\in C_{t}^{m}}N_{c(i)}(\widehat{\varepsilon}_{it})}{\sum_{i\in C_{t}^{m}}N_{c(i)}}\right)k_{mt}$$

By Assumptions 1-2:

$$\mathbb{E}\left(\sum_{t}w_{mt}\frac{\sum_{i\in C_{t}^{m}}N_{c(i)}(\widehat{\varepsilon}_{it})}{\sum_{i\in C_{t}^{m}}N_{c(i)}}\right)k_{mt}=$$
$$\mathbb{E}\left(\sum_{t}w_{mt}\frac{\sum_{i\in C_{t}^{m}}N_{c(i)}((\lambda_{it}^{M}-\lambda_{it}^{L})+(\zeta_{c(i)}^{M}-\zeta_{c(i)}^{L})+(\omega_{mt}-\omega_{lt}))}{\sum_{i\in C_{t}^{m}}N_{c(i)}}\right)k_{mt}=0$$

Therefore, it follows that:

$$\mathbb{E}[\widehat{TVA}_{m}]=\mathbb{E}\left(\sum_{t}w_{mt}k_{mt}\widehat{\mu}_{m}\right)+\mathbb{E}\left(\sum_{t}w_{mt}\frac{\sum_{i\in C_{t}^{m}}N_{c(i)}(\widehat{\varepsilon}_{it})}{\sum_{i\in C_{t}^{m}}N_{c(i)}}\right)k_{mt}$$
$$=\sum_{t}w_{mt}k_{mt}\overline{TVA}_{m}$$

For any shrinkage factor $\sum_{t}w_{mt}k_{mt}\neq 1$, it follows that $\widehat{TVA}_{m}$ is not unbiased, but is instead forecast unbiased.

A.5 Spillover effects

Note that, following the literature, the assumed production function of a student's test score assumes no spillover teacher effects across subjects. In other words, a student's Literature teacher does not influence her Math score and vice versa. Given the use of cross-sectional variation in scores, we need to formally discuss this assumption and the implications of non-zero spillover.

For this purpose, let us assume there is instead a positive teacher spillover across subjects that is a linear function of a teacher's TVA in their own subject:

$$\overline{TVA}_m^L = r^M + q^M \overline{TVA}_m \text{ and } \overline{TVA}_l^M = r^L + q^L \overline{TVA}_l \text{ where } q^M > 0 \text{ and } q^L > 0$$

It follows that we need to adjust the sum of the underlying unobserved factors as:

$$\begin{aligned}
\widehat{A}_{it} =& \overline{TVA}_m - \overline{TVA}_l + (\lambda_{it}^M - \lambda_{it}^L) + (\zeta_{c(i)}^M - \zeta_{c(i)}^L) + (\omega_{mt} - \omega_{lt}) \\
&+ (r^L - r^M) + (q^L \overline{TVA}_l - q^M \overline{TVA}_m) \\
&= (1 - q^M)\overline{TVA}_m - (1 - q^L)\overline{TVA}_l + (\lambda_{it}^M \lambda_{it}^L) + (\zeta_{c(i)}^M - \zeta_{c(i)}^L) + \\
&- (\omega_{mt} - \omega_{lt}) + (r^M - r^L)
\end{aligned}$$

As $q^M > 0$, both $\widehat{A}_{mt}$ and the subsequent $\widehat{\overline{TVA}}_m$ the estimated difference of TVA will be biased towards zero by the amount $(1 - q^M)$.

ASSUMPTION A.3. (Spillover effects): *There are no spillover effects from a teacher in a network to a student's score in the subject of the link teacher:*

$$\overline{TVA}_m^L = 0$$

The assumption entails that the choice of subjects to identify TVA leads to a trade-off. On the one hand, the closer the subjects are in skills, the larger the share of common ability would be compared to subject-specific ability. A good example is the use of Math and Physics: the two subjects require many of the same skills, such that a larger share of ability necessary for the Math and Physics exams would be common between the two subjects, and would therefore be canceled out. On the other hand, the closer the two subjects are in required skills, the more likely that there are spillovers from one teacher to the other's subject. It is plausible to imagine that a good Math teacher can positively contribute to a student's Physics score.

The assumption is more plausible when using Math and Literature teachers, as Math and Literature are two subjects with some skill overlap (e.g. logical thinking), but no likely spillovers between teachers at the middle school level, even if spillovers are possible at earlier stages of the education process.

B Simulation setup and results

B.1 Setup

The simulations are based on a school with the properties of the representative school in the French data in terms of the total number of classrooms, the number of students per class, and number of teachers, over a period of 10 years (see Table B.2). In what follows, I focus on simulations for the value added of Math teachers, but naturally, the equivalent is true for the value added of French teachers.

Table B.1: Simulated school

Note: The table depicts the characteristics of the average school in my sample of schools for a total period of 10 years. These characteristics are used for the purpose of simulating the average school.

Characteristic	**Size**
Total number of schools	1000
Avg. number of classrooms per school and year	3.5
Total number of classrooms per school (over 10 yrs)	40
Avg. classroom size	24
Classroom size range	[15,30]
Avg. number of z teachers per school and year	8
Total number of z teachers per school (over 10 yrs)	8
Avg. number of classrooms per teacher and year	1.3
Range of number of classrooms per teacher (over 10 yrs)	[2,6]

B.2 Data Generating Process

Each student is observed only in only two periods, $t-1$ and t, such that sorting at $t-1$ is random, but sorting at t may not be.

Student i's score at period $t-1$ is determined by:

$$A^z_{i,t-1} = \lambda^z_{i,t-1} + \lambda_{i,t-1} + \mu_{j(z,t-1)} + \zeta^z_{c,t-1} + \epsilon_{i,t-1} + \epsilon^z_{i,t-1} \tag{B.6}$$

where $\lambda^z_{i,t-1}$ is $t-1$ period's ability of student i in subject z, $\lambda_{i,t-1}$ is i's common ability, $\mu_{j(z,t-1)}$ is the TVA of the teacher in subject z, $\zeta^z_{c,t-1}$ are classroom-level c's idiosyncratic effects and $\epsilon_{i,t-1}$ is an additional student and time-specific idiosyncratic noise. One can think of $\zeta^z_{c,t-1}$ as a noise which affects all the students in classroom c in the same way, with a constant variation of the noise across years (i.e. the same noise would affect students in grade 8th and 9th in the same way) but no autocorrelation across years (e.g. the exam taking place in a noisy environment at $t-1$ due to works on the street does not affect exam results at t in any way). We can think of the student noise $\epsilon_{i,t-1}$ in a similar way: being ill on the

day of the exam would affect the grade of student i in grade 8th and 9th in the same way, but if i was sick at $t-1$, this would not affect her score at t.

Denote the persistence of ability in period t as η. It follows that in period t, the score of i is:

$$A^z_{it} = \eta\lambda^z_{i,t-1} + (1-\eta)\lambda^z_{it} + \eta\lambda_{i,t-1} + (1-\eta)\lambda_{it} + \mu_{j(z,t)} + \zeta^z_{ct} + \epsilon_{it} + \epsilon^z_{it} \tag{B.7}$$

In other words, ability of student i at time t is a function of i's ability at $t-1$ (with a factor of persistence η), and some new ability factor, discounted with a factor of $(1-\eta)$, to keep the variance of each of the ability factors unchanged.

I follow Chetty et al. (2014a)'s findings concerning the SD of student, classroom and teacher effects as parameters in my simulation (see Table B.2). I make additional assumptions about the shares of the variance of each student component. Specifically, the student-level variance is split into the variances of subject-specific ability, common ability and idiosyncratic noise of the transmission from ability to exam score. I assume that the share of common ability in the variance is larger that of the subject-specific ability share. The noise in the prediction of exam scores is added in two ways, as shown in equation B.2: through the additional gains in ability at time t, $(1-\eta)\lambda^z_{it}$ and $(1-\eta)\lambda_{it}$, and through the idiosyncratic student noise ϵ_{it}.

Table B.2: Exogenous simulation parameters

Note: The table depicts the inputs in the simulations used to derive student test scores in Math and Literature. The standard deviations of TVA and class effect correspond to the results of Chetty et al. (2014a) for test scores in Math (here extended to both Math and Literature for tractability and to reduce the necessary assumptions needed for analysis). The variance of student noise is split into variances of subject-specific and common ability, as well as random student noise. I assume the share of common ability in the variance is close to double that of the subject-specific ability. The noise in the prediction of exam scores is added in two ways, as shown in equation B.2: through the additional gains in ability at time t, $(1-\eta)\lambda^z_{it}$ and $(1-\eta)\lambda_{it}$, and through the idiosyncratic student noise ϵ_{it}.

Simulation parameter	**Mean**	**SD**
TVA	0	0.134
Class effect	0	0.116
Subject-specific ability	0	0.455
Common ability	0	0.729
Persistence of ability	0.8	0
Idiosyncratic noise	0	0.595

B.3 Simulating non-random sorting

Three out of the four simulation scenarios are based on settings with non-random sorting, based on a specific sorting variable: (i) the average lagged score between Math and Literature,

(ii) the difference between the lagged Math and Literature scores, and (iii) the change in student motivation.

I simulate non-random sorting in the following way. First, I sort students to classrooms based on their sorting variable value. For example, if the sorting variable is the average lagged score, I compute the average score for each student in period $t-1$ and place each student in period t into a new class based on their average lagged score, such that students with the highest lagged scores are grouped together in a class (given the pre-defined restriction of class size), students with marginally worse scores are grouped together in another class, and so on.

After placing students to classes, I allocate Math teachers to classes, using pre-defined restrictions on the number of classes each teacher teaches and the specific years in which the teacher teaches. Specifically, the highest TVA Math teacher is placed in front of the class(es) with the highest values of the sorting variable, the second highest TVA Math teacher is placed in front of the class(es) with the marginally lower values of the sorting variable, and so on. I allow Literature teachers to be placed randomly across classes. Alternative specifications which sort both teachers non-randomly lead to similar but noisier results, as the very strict sorting pattern allows for much fewer connected sets (the best Math teachers are observed only with the best Literature teachers, and vice versa).

B.4 Results

The sorting on lagged average score case This section shows the results corresponding to the scenario of non-random sorting of students to classrooms based on their lagged average score in Math and Literature, and positive assignment of Math and Literature teachers to classrooms (i.e. better teachers assigned to classrooms which are better on average).

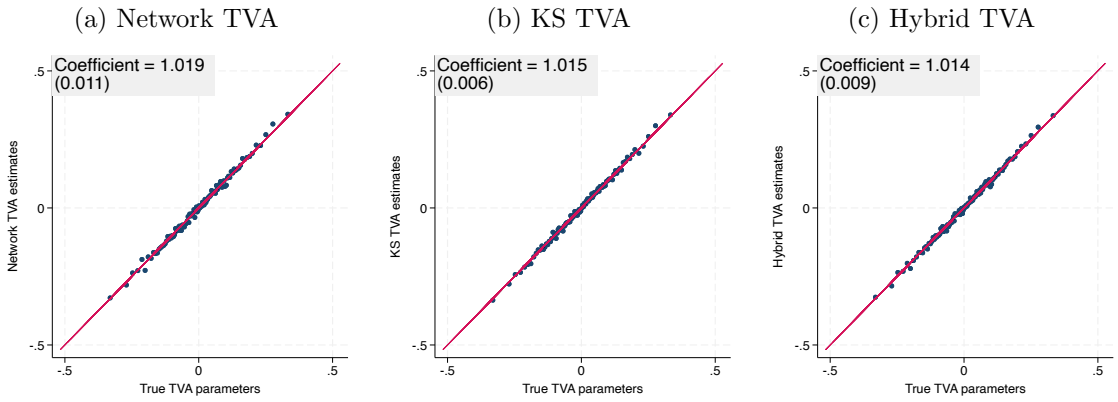


Figure B.2: Relationship between estimates of TVA and true TVA parameters

Note: The figures are based on the analysis where sorting of teachers to students is based on lagged average scores. Each binscatter represents the relationship between the TVA estimates (on the y-axis) and the true TVA parameters (on the x-axis) within school, such that Panel (a) reports the results for the Network estimator, Panel (b) reports the results for the KS estimator, and Panel (c) reports the results for the Hybrid estimator. The TVA parameters are divided into 100 equal-sized groups and each dot represents the mean value of $\widehat{A}_m$ in each bin. The red line is the 45-degree line. The gray boxes in each panel reports the regression coefficient and standard error from a regression of $\widehat{A}_m$ on μ_m, controlling for school fixed effects, rather than the binned averages. Standard errors are clustered at the school level.

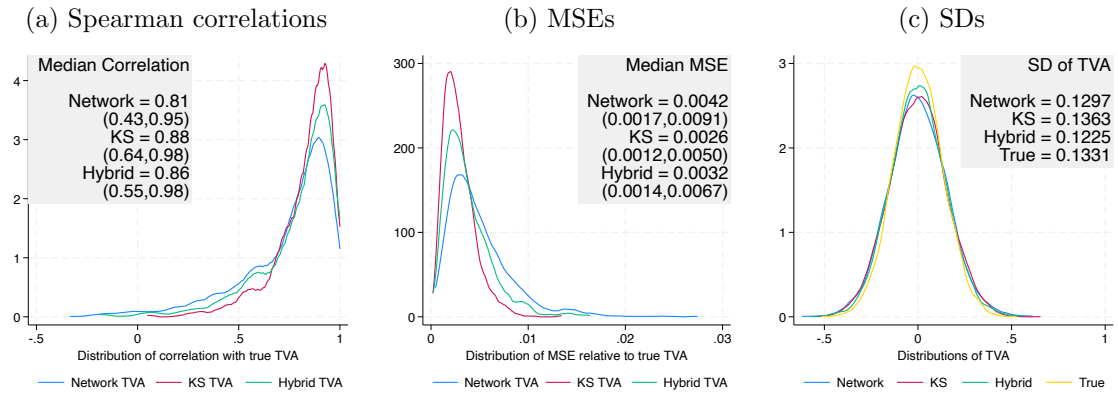


Figure B.4: Predictive power of estimates compared to true TVA

Note: The figures are based on the analysis where sorting of teachers to students is random. Panel (a) depicts the distribution of Spearman correlations between the Network, Hybrid or KS estimates of TVA and the true TVA parameters respectively, computed at the school level. The gray box shows the correlation for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (b) depicts the distribution of mean squared errors (MSEs) of the Network, Hybrid or KS estimates of TVA compared to the true TVA parameters respectively, averaged at the school level. The gray box shows the MSE for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (c) shows the empirical distributions of the TVA estimators and the distribution of true TVA. The gray box shows the covariance of yearly teacher residuals for each method and the SD of true TVA.

Sorting on the difference between lagged Math and Literature scores and positive teacher assignment This section shows the results corresponding to the scenario of non-random sorting of students to classrooms based on the difference between their lagged Math and Literature scores, and positive assignment of Math teachers to classrooms (i.e. better teachers assigned to classrooms which are better on average).

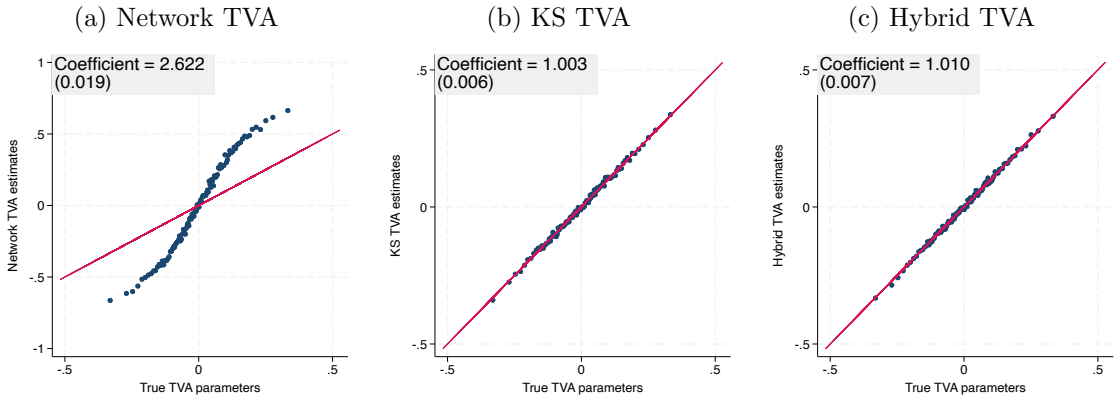


Figure B.6: Relationship between estimates of TVA and true TVA parameters

Note: The figures are based on the analysis where sorting of teachers to students is based on the difference between their lagged Math and Literature scores. Each binscatter represents the relationship between the TVA estimates (on the y-axis) and the true TVA parameters (on the x-axis) within school, such that Panel (a) reports the results for the Network estimator, Panel (b) reports the results for the KS estimator, and Panel (c) reports the results for the Hybrid estimator. The TVA parameters are divided into 100 equal-sized groups and each dot represents the mean value of $\widehat{A}_m$ in each bin. The red line is the 45-degree line. The gray boxes in each panel reports the regression coefficient and standard error from a regression of $\widehat{A}_m$ on μ_m, controlling for school fixed effects, rather than the binned averages. Standard errors are clustered at the school level.

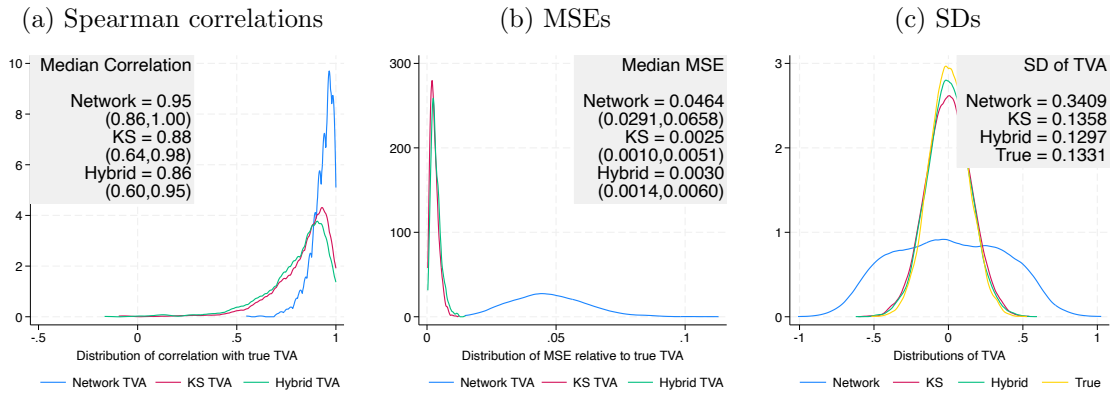


Figure B.8: Predictive power of estimates compared to true TVA

Note: The figures are based on the analysis where sorting of teachers to students is based on the difference between their lagged Math and Literature scores. Panel (a) depicts the distribution of Spearman correlations between the Network, Hybrid or KS estimates of TVA and the true TVA parameters respectively, computed at the school level. The gray box shows the correlation for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (b) depicts the distribution of mean squared errors (MSEs) of the Network, Hybrid or KS estimates of TVA compared to the true TVA parameters respectively, averaged at the school level. The gray box shows the MSE for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (c) shows the empirical distributions of the TVA estimators and the distribution of true TVA. The gray box shows the covariance of yearly teacher residuals for each method and the SD of true TVA.

Sorting on the change in students' motivation and positive teacher assignment

This section shows the results corresponding to the scenario of non-random sorting of students to classrooms based on the change in their motivation, and positive assignment of Math teachers to classrooms (i.e. better teachers assigned to classrooms which have become more motivated on average).

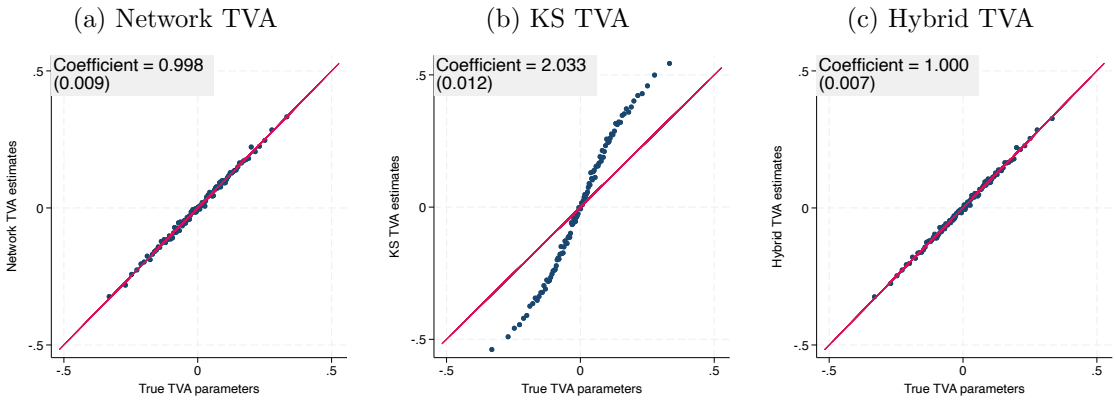


Figure B.10: Relationship between estimates of TVA and true TVA parameters

Note: The figures are based on the analysis where sorting of teachers to students is based on the change in students' motivation. Each binscatter represents the relationship between the TVA estimates (on the y-axis) and the true TVA parameters (on the x-axis) within school, such that Panel (a) reports the results for the Network estimator, Panel (b) reports the results for the KS estimator, and Panel (c) reports the results for the Hybrid estimator. The TVA parameters are divided into 100 equal-sized groups and each dot represents the mean value of $\widehat{A}_m$ in each bin. The red line is the 45-degree line. The gray boxes in each panel reports the regression coefficient and standard error from a regression of $\widehat{A}_m$ on μ_m, controlling for school fixed effects, rather than the binned averages. Standard errors are clustered at the school level.

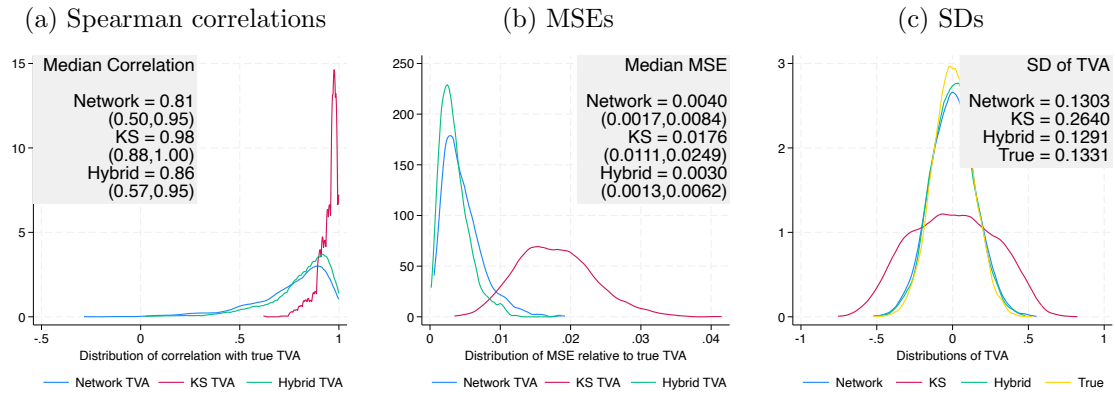


Figure B.12: Predictive power of estimates compared to true TVA

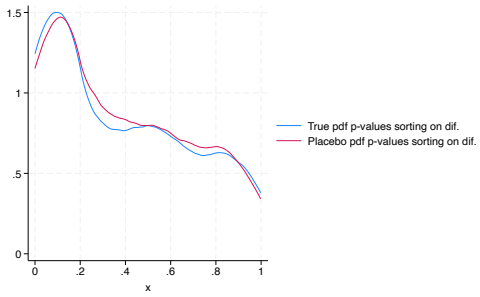
Note: The figures are based on the analysis where sorting of teachers to students is based on the change in students' motivation. Panel (a) depicts the distribution of Spearman correlations between the Network, Hybrid or KS estimates of TVA and the true TVA parameters respectively, computed at the school level. The gray box shows the correlation for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (b) depicts the distribution of mean squared errors (MSEs) of the Network, Hybrid or KS estimates of TVA compared to the true TVA parameters respectively, averaged at the school level. The gray box shows the MSE for the median school for each estimate with the true parameters respectively, and the interval of the 10th and 90th percentile. More weight is given to schools with more students. Panel (c) shows the empirical distributions of the TVA estimators and the distribution of true TVA. The gray box shows the covariance of yearly teacher residuals for each method and the SD of true TVA.

C Graphs of tests of sorting

Figure C.14: PDF of p-values for difference

Note: Distribution of F-test p-values within school within year for real assignment vs. simulated assignment for the sample of teachers with more than 1 year within school. The type of sorting refers to the way in which the placebo re-assignment of teachers is conducted. Type 1 sorting takes only 8th grade teachers at a particular year and reallocates them to 8th grade classrooms within the school. Type 2 sorting takes all middle school teachers at a particular year and reallocates them across classrooms within the school. I then take only 8th grade classrooms for the analysis.

(a) Type 1 sorting



(b) Type 2 sorting

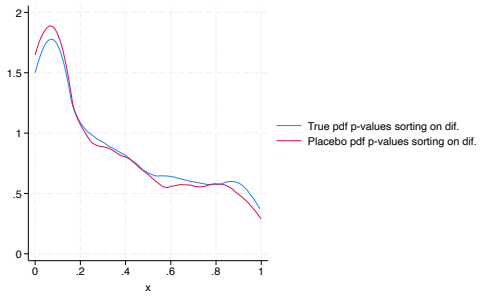
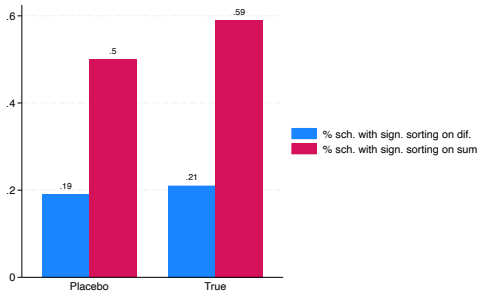


Figure C.15: Grade 8: Perc. of schools with p-value <= 0.05

Note: Perc. of schools with p-value <= 0.05 for sorting within school within year for real assignment vs. simulated assignment for the sample of teachers with more than 1 year within school. The type of sorting refers to the way in which the placebo re-assignment of teachers is conducted. Type 1 sorting takes only 8th grade teachers at a particular year and reallocates them to 8th grade classrooms within the school. Type 2 sorting takes all middle school teachers at a particular year and reallocates them across classrooms within the school. I then take only 8th grade classrooms for the analysis.

(a) Type 1 sorting



(b) Type 2 sorting