Teacher Value-Added in the Absence of Annual Test Scores: Utilising Teacher Networks

Latest Version

Desislava Tartova [*]

June 19, 2023

Abstract

I develop a novel method for estimating teacher value added which controls for non-random student-teacher sorting without having to control for lagged grades in standardised tests. To do so, I exploit "networks" of teachers - teachers from the same subject who are observed in classrooms with a unique "link" teacher from another subject. I measure the relative value added of two teachers in a network as the difference between their classrooms' grades in a standardised exam, unexplained by student characteristics, correcting for the classrooms' grade differential in the subject of the link teacher. I show that the estimated teacher effects are unbiased under plausible assumptions that I confirm in the data. Using exhaustive French administrative data, I find that a 1 SD increase in teacher value added within school improves student scores by 0.17 SD in Math and 0.16 SD in French.

[*]Paris School of Economics (PSE). Email address: desislava.tartova@psemail.eu. I am very grateful to Luc Behaghel for his guidance and support. I also thank Barbara Biasi, Julien Grenet, Nina Guyon, Antoine Hubert de Fraisse, Adrien Matray, Adam Kapor, Sofia Betania Sierra Vasquez, Camille Terrier, Yves Zenou, as well as participants at the CESifo/ifo Junior Workshop on the Economics of Education, Doctorissimes conference, SSE Quality in Education conference, PSE Applied Economics seminar, PSE Labour and Public Economics seminar and the ENS Workshop in Economics of Education. 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.

1 Introduction

Teacher value added - the effect of a teacher on her students' grades - has been consistently identified as a better predictor of student long-term outcomes than any observable teacher characteristics.[1] Nevertheless, the majority of educational systems around the world rely heavily on observable characteristics such as the teacher's exam grades, qualification and experience in order to make decisions on recruitment, compensation and career progression, and do not take into account measures of teacher value added.

One important reason for this fact is that standard methods for value-added estimation (e.g. Rockoff, 2004; Kane & Staiger, 2008; Chetty et al., 2014a) rely on a panel data of student test scores in standardised exams, which is missing in most countries due to the specificities of their educational systems. For instance, in the European Union, only two countries conduct standardised assessments in two consecutive years over the course of a student's academic experience.[2]

The panel structure of the data is important because failing to control for lagged exam scores may lead to bias in the teacher value added estimates. More precisely, to identify the value added of teachers, researchers regress student test scores on a function of lagged test scores, other student characteristics and teacher fixed effects - such that the latter identify a teacher's value added. Importantly, lagged test scores are assumed to be a sufficient statistic to control for unobserved student characteristics, such as ability. Hence, including lagged test scores would control for non-random sorting of teachers across classrooms based on student unobservables: for instance, high-ability students being placed in classrooms taught by high-quality teachers. Omitting past grades in such case would lead to an overestimation of the value added of these teachers.

This paper develops a method to uncover plausibly unbiased relative teacher value added estimates within school in settings where such panel data analyses are not possible, relying only on cross-sectional student information in more than one subject. In simulated settings, I show that the method performs well compared to the standard method under minimal assumptions. I confirm the plausibility of these assumptions using French administrative data covering the universe of public middle schools in France between 2009 and 2018, and taking advantage of the standardised tests in Math and French performed at the end of the 9th grade. Applying the method to the French setting, I find that for a 1 standard deviation increase in the value added of a Math (French) teacher within a school, the average student's test score improves by 17.4 (16.3) percent of a standard deviation.

To grasp the intuition behind the method, it is worth starting from a general approach.

[1] For instance, Rockoff (2004), Kane, Rockoff, and Staiger (2008), Chetty, Friedman, and Rockoff (2014a), and Chetty, Friedman, and Rockoff (2014b), Bau and Das (2020).

[2] One notable exception is the United States, on which the bulk of the literature has focused thus far.

One can think of a simplified case of a school with only two classrooms, each taught by a different Math teacher and a different French teacher. Let us assume that observable student characteristics do not matter for student grades and only teacher value added and unobservable student ability explain grades. We can take the difference between the average grades in Math between both classrooms (denoted *difference M*). This difference measures the sum of both the differences in value added between the two Math teachers and the difference in average student ability between the two classrooms. By analogy, the same difference between French grades of both classrooms (denoted *difference F*) measure the sum of both the differences in value added between the two French teachers and the difference in average student ability between the two classrooms. By taking the difference between *difference M* and *difference F*, we obtain a measure of the difference between the difference in value added between the two Math teachers and the difference in value added between the two French teachers.

My method focuses on a corner case of this general approach in which the two Math teachers are observed with a single French teacher. As the value added of the French teacher in both classrooms is by construction the same, the difference between *difference M* and *difference F* represents the relative value added between the two Math teachers.

More specifically, I propose to uncover relative value added of two teachers by exploiting such differences in differences within "networks" of teachers: teachers in the same subject who have been observed in a classroom with the same teacher from another subject - a "link" teacher. I measure the relative value added of two such teachers in a network in a two-step procedure. In the first step, I residualise student grades in standardised exams in Math and French from observable student characteristics, such as parental socio-economic status, the scholarship status of the student, gender, age, and region of birth. In the second step, I compute the difference in the average residualised Math grades of the two classrooms, controlling for the average overall level of ability by subtracting the difference in the two classrooms' residualised French grades.

Transforming such pairwise network comparisons into a distribution of teacher value added which includes all teachers within a school is achieved by transitivity. For the purpose of example, imagine a school with three Math teachers in total - A, B and C. Teachers A and B form a "network", as do teachers B and C, but teachers A and C are not observed together. Using the two available relative value added estimates in a system of equations allows me to find the relative value added of teachers A and C. I define a school in which I can identify the entire value added distribution through either direct or indirect networks a "complete school network", and focus exclusively on such schools.

By construction, the estimation method allows to control for any potential sorting of teachers to students based on the overall ability - the ability that is common between the two subjects and would therefore be captured when controlling for the grade in the subject of the link teacher. A problem may arise if there exist students that have subject-specific ability which deviates from their overall ability, and in addition teachers are sorted across classrooms based on that subject-specific ability. For example, if high-quality Math teachers are allocated to classrooms which are on average relatively better in Math than they are in French, then one would overestimate their value added.

It follows that the unbiasedness of the value added estimates relies on the assumption that Math (French) teachers are not sorted to classrooms based on the relative Math-specific (French-specific) ability in these classrooms. While the assumption cannot be tested for unobservable student characteristics such as ability, I can test whether there seems to be sorting between teachers and students based on observable characteristics. To the extent that the observable characteristics are correlated with the unobserved ability, such tests can give us an idea of the sorting patterns on ability as well.

To test this assumption, I first predict the student-level difference between her Math and French grades with observable student characteristics, such as the socio-economic status of her parents and her scholarship status. For every school, I then regress the predicted difference in grades on teacher fixed effects. I compare the average R^2 of such regressions to the distribution of R^2 obtained with bootstrapped data under the null of no sorting based on the difference in Math and French grades predicted by observable characteristics. I find that the probability of such sorting is lower or equal to the one obtained in the simulated random assignments, suggesting that I fail to reject the null of no such sorting. This finding gives supporting evidence to the identifying assumption.

In addition to the main identifying assumption, I derive two more assumptions necessary for identification of the value-added estimates. In particular, I assume that each classroom in which a teacher is observed is equally informative of the teacher's value added. I conduct several tests which attest to that. Moreover, I assume that any time component (or "drift", as in Chetty et al. (2014a)) of value added is additively separable from the intrinsic value added of a teacher and is a function of the years of experience of the teacher, and I control for it by including years-of-experience fixed effects in the first stage of the analysis.

I then compare my method to the standard method in a simulated school setting that matches the characteristics of the average French middle school in my data. I conduct three types of simulations exercises. I focus on a case with random sorting between teachers and students, a case with sorting based on the overall ability and a case with sorting based on subject-specific ability. I evaluate the mean squared error (MSE) and Spearman correlation (SP) of both the proposed method and the standard method compared to the true value added parameters averaging over 1,000 simulations. In the case of no sorting, my method performs almost equally well in terms of MSE, producing a very small MSE compared to the standard deviation of the value added parameters. Once sorting on common ability is introduced, consistent with the theoretical prediction of my method, the MSE is unchanged. On the other hand, the square root of the MSE of the traditional method is almost 50 times as big as the one of my method. Finally, consistent with the predictions, once sorting on subject-specific ability is introduced, the MSE of my method drastically increases. It is, however, only 2 times as big as the MSE of the traditional method. Finally, in all three scenarios, both methods produce perfect SP.

I test the method empirically using French administrative data for the universe of middle schools in Metropolitan France. Focusing on 9th graders and their Math and French language teachers, I find that for a 1 standard deviation increase in Math (French) teacher value-added within a school, student scores improve by 17.4 (16.3) percent of a standard deviation. This implies that moving a student from a teacher who is at the 5th percentile of the value-added distribution of a school to one who is at the 95th percentile of the same value-added distribution would increase that student's test scores on average by 58 percent of a standard deviation in Math and by 54 percent of a standard deviation in French. I show that the variation of value added is higher in more disadvantaged regions of France and schools which have a higher share of disadvantaged students. In addition, I show that the teacher estimates are positively correlated with being female and with in-class pedagogical assessments, and are not associated with having a higher salary or an advanced qualification, consistent with the literature.

The paper's main contribution is to expand the scope of the teacher value added literature to educational systems where students take standardised exams in more than one subject in a single year, but do not necessarily take exams in the same subject in two consecutive years. Cross-sectional student grade data in different subjects has been used in a number of settings so far, from studies which focus on the effect of instructional time (Lavy, 2015), to such which link teacher credentials or teacher assessment grades to student achievement (e.g. Clotfelter, Ladd, & Vigdor, 2010; Benhenda, 2018). Nevertheless, it has so far not been used in the setting of value-added estimation.

I also contribute to the literature by providing empirical evidence on teacher value-added non-experimentally outside of the widely studied United States setting or developing countries. Indeed, these results provide the first test of value added in French middle schools and one of the first non-experimental evidence in Europe. The estimates of within-school value added in France are slightly larger compared to those in United States middle schools, which typically lie between 0.10 and 0.15 SD in Math and 0.05 and 0.15 SD in Literature (Jackson, 2014,Bacher-Hicks & Koedel, 2022), and are more in line with estimates found for some developing countries (e.g. Bau & Das, 2020; Buhl-Wiggers, Kerwin, Smith, & Thornton, 2017).[3] They are also in line with the typical finding that Math teachers have a higher variability in value added compared to Literature teachers (Lefgren & Sims, 2012; Condie, Lefgren, & Sims, 2014,Bau & Das, 2020). This gives further evidence that the effect of teachers on student outcomes may be institution-specific, stemming from the differences in the educational environment across countries.

The rest of the paper proceeds as follows. Section 2 briefly discusses the related literature. Section 3 derives the proposed methodology and the necessary identification assumptions. Section 4 outlines the corresponding empirical strategy, whereas Section 5 provides a simulation exercise. Section 6 details the French institutional setting, the data sources and descriptive statistics. Section 7 provides the results. Finally, section 8 concludes.

2 Teacher value added in the literature

The existing literature on value-added modelling has not changed substantially since its creation (see Bacher-Hicks & Koedel, 2022 for a recent extensive review). Although there is variation in the exact choice of regression and the set of covariates in the literature, the essential components across all specifications are similar and rely on the same underlying theoretical model. From a theoretical perspective, one can decompose the student grade of student i in subject f into a set of factors,

$$A_{i,f,s,t} = A[\mathbf{X}_i, \gamma_{i,f}, \theta_{j_f,t}, \theta_s, \epsilon_{i,f,s,t}]$$

or, assuming additive separability of inputs, as is common in the teacher value-added literature, we can re-write the above equation as:

$$A_{i,f,s,t} = \mathbf{X}_i\beta_f + \gamma_{i,f} + \theta_{j_f,t} + \theta_s + \epsilon_{i,f,s,t} \tag{1}$$

In this simple function, $A_{i,f,s,t}$ is the grade of student i in subject f, and $\theta_{j_f,t}$ represents the value added of i's teacher j_f at time t. One can think of the teacher value added $\theta_{j_f,t}$ as a function $\theta_{j_f,t} = \theta_{j_f} + \alpha_{j_f,t}$ - a combination of a permanent component of teacher quality θ_{j_f}, and a time-specific fluctuations of teacher value added (or a "drift" of value added, as referred to by Chetty et al. (2014a)), $\alpha_{j_f,t}$. In addition, X_i is a vector of other observable student characteristics, $\gamma_{i,f}$ is the underlying ability of student i in subject f, θ_s represents school productivity, and $\epsilon_{i,f,s,t}$ is an idiosyncratic error which contains all other possible inputs that impact i's grade in subject f.

Importantly, an essential part of teacher value added models and indeed a common component across all value added papers is the identification of the student's unobserved characteristics, $\gamma_{i,f}$. We can think of $\gamma_{i,f}$ as some combination of average ability and subject-specific

[3] Bau and Das (2020) finds a SD of 0.19 for Math teachers, 0.11 for Urdu teachers and 0.15 for English teachers in Pakistan. Buhl-Wiggers et al. (2017) finds an average SD across subjects between 0.14 and 0.16.

ability, $\gamma_{i,f} = \theta_i + \theta_{i,f}$. The majority of papers proxy for $\gamma_{i,f}$ empirically by including $A_{i,f,t-1}$ in their regression analysis - student i's grade in the subject in the previous academic year. In particular, student i's grade in the subject from the previous year is seen as a sufficient statistic to account for the student's unobserved ability endowment in subject f and all other factors, such as parental and school inputs, that may affect i's performance in period t and are not included in $\mathbf{X}_i$.

Including such lagged grades is important, as students might be sorted to teachers in a non-random manner based on the students' ability, which would bias teacher value added estimates if not controlled for.[4]

In particular, if higher-quality teachers within a school teach only the classes with higher-ability students (and vice versa for lower-quality teachers), estimating equation 1 without past grades would lead to an upward bias in estimates - one would overestimate the quality of higher-quality teachers and underestimate the quality of lower-quality teachers. A number of papers assess whether simply controlling for past grades is enough to obtain unbiased estimates (a notable example is Chetty et al., 2014a) and typically conclude that in the presence of non-random sorting, the model tends to recover reasonable estimates, albeit slightly biased.

However, due to differing cultural and political preferences, many countries do not conduct standardised assessments on students in consecutive years of the students' academic path. Strikingly, in the European Union, only two countries conduct at least two standardised tests over two consecutive years: France in elementary school and Malta in middle school.[5] This automatically means that the method is inapplicable for the majority of European countries, as well as for the majority of other countries, explaining the strikingly small number of studies of value added outside of the US. This stylised fact motivates the method proposed in this paper.

[4]It is worth noting that the term is not necessary if students are randomly allocated to classrooms and randomly matched to teachers. If this is the case, any variation in student ability across classrooms would on average add random noise to the estimates of teacher value added rather than bias.

[5]More specifically, these are France and Malta. Since September 2018, French pupils in the 1st and 2nd grade (CP and CE1, respectively) undergo national assessments ("Repères") in Mathematics and French language. In the 1st grade, these assessments are administered twice during the school year - in September and February, whereas in the 2nd grade, there is only one assessment in September. The next possible standardised assessment is only held in the 6th grade. By contrast, in Malta standardised assessments are conducted at the end of 4th, 5th and 6th grade in Maltese, English and Maths. Two more examinations are held in high school, at the age of 16 and 18.

3 Proposed value-added methodology

In this section, I propose and discuss in detail the identification of teacher value-added with the use of within-student across-subject variation in test scores. I focus on a setting where one cannot proxy student i's unobservable factors in subject f, denoted $\gamma_{i,f}$, by the lagged grade student i grade in subject f, $A_{i,f,t-1}$. However, there are strictly more than one subjects assessed in year t, such that the two different subjects are taught to a student by two different teachers. For simplicity, and so as to mirror the actual strategy used for the empirical test, I focus on two subjects in my notation - Math and French, such that $f \in \{M, F\}$.[6]

3.1 Simplified case of two classrooms

I propose to make use of the cross-sectional variation in grades in different subjects of students in order to uncover the distributions of teacher effects in each subject.[7] For tractability, I average student grades on a classroom-subject level, and I first assume a school with only two classrooms, $c \in \{c_A, c_B\}$ of equal size that are observed at the same time t. To simplify notation, I omit the subscript t from the equations of interest for the time being.

Intuition using two teachers in each subject Assume that each classroom has one Math teacher, $j_M \in \{j_{M_A}, j_{M_B}\}$ and one French teacher, $j_F \in \{j_{F_A}, j_{F_B}\}$, such that j_{M_A} and j_{F_A} teach at classroom c_A, and j_{M_B} and j_{F_B} teach at classroom c_B. For each classroom in each subject f, I can express the average classroom grade as a function of the same characteristics as:

$$\overline{A}_{c,f} = \overline{\mathbf{X}}_c \beta_f + \gamma_{c,f} + \theta_{j_f} + \theta_s + \varepsilon_{c,f} \tag{2}$$

where θ_{j_f} is the teacher's effect on the grade of the student, θ_s are school-related effects, for example due to peer effects and the effect of the principal, $\overline{\mathbf{X}}_c$ represent observable student characteristics, and $\gamma_{c,f}$ represents the unobserved classroom characteristics in a subject, which we can think of as the average unobserved characteristics of students in classroom c.

[6]Following Benhenda (2018), the choice to restrict my attention on Math and French is done for two main reasons. Firstly, this is useful for the purpose of comparison with the rest of the literature, which focuses exclusively on Math and Reading/Literature/English (Jackson, 2014). Secondly, Math and French are two subjects in which the threat of teacher value-added spillover across subjects is minimal (Koedel, 2009). By contrast, one would expect that spillover effects seem more plausible between History-Geography and French, or Math and Physics-Chemistry.

[7]Note that simply adding student fixed effects would not help with identification. While estimates would be unbiased provided that teacher-student sorting is based on general, rather than subject-specific ability, we would essentially be comparing Math to French teachers, rather than Math to Math (and French to French) teachers. This is not desirable, as according to the existing literature, teacher effects vary significantly across subjects (Lefgren & Sims, 2012; Condie et al., 2014) and are thus not directly comparable.

I can decompose $\gamma_{c,f}$ for Math and French into the average common ability component in the classroom and the average subject-specific ability, such that $\gamma_{c,f} = \theta_c + \theta_{c,f}$.

To better understand the identification, note that for each classroom, I can express equation 2 in the equivalent difference in grades across subjects, such that for classroom c_A:

$$\overline{A}_{c_A,M} - \overline{A}_{c_A,F} = \overline{\mathbf{X}}_{c_A} \cdot (\beta_M - \beta_F) + (\theta_{j_{M_A}} - \theta_{j_{F_A}}) + (\theta_{c_A,M} - \theta_{c_A,F}) + (\varepsilon_{c_A,M} - \varepsilon_{c_A,F})$$

and equivalently, for classroom B. Subtracting the equations for classroom c_B from that for c_A and rearranging leaves us with a difference-in-difference equation of the sort:

$$\Delta A_{M,c_A,c_B} = (\overline{\mathbf{X}}_{c_A} - \overline{\mathbf{X}}_{c_B}) \cdot (\beta_M - \beta_F) + (\theta_{j_{M_A}} - \theta_{j_{M_B}}) - (\theta_{j_{F_A}} - \theta_{j_{F_B}}) + \eta_{c_A,c_B,M,F} \quad (3)$$

where

$$(\theta_{c_A,M} - \theta_{c_B,M}) - (\theta_{c_A,F} - \theta_{c_B,F}) + (\varepsilon_{c_A,M} - \varepsilon_{c_B,M}) - (\varepsilon_{c_A,F} - \varepsilon_{c_B,F}) \equiv \eta_{c_A,c_B,M,F}$$

and

$$(\overline{A}_{c_A,M} - \overline{A}_{c_A,F}) - (\overline{A}_{c_B,M} - \overline{A}_{c_B,F}) \equiv \Delta A_{M,c_A,c_B}$$

Note that in this setting, one can directly now compare Math to Math teachers and French to French teachers. In particular, we can find the relative value-added of the Math teachers j_{M_A} and j_{M_B}, controlling for the difference in value-added of each classroom's French teacher (j_{F_A} and j_{F_B}, respectively). However, what this means is that in addition to capturing the difference in value added of the Math teachers, we also capture the difference in value added of the French teachers.

In the best case, this would lead to noise in our estimates of Math teachers' relative value added. This is the case as the estimated value added of the Math teachers would be different if measured by controlling for a different set of French teachers, for an unchanged true relative value added of the two Math teachers. To see this, imagine a simple scenario where the true value-added of j_{M_A} is higher than the true value added of j_{M_B}. If j_{M_A} happened to randomly be coupled with a very high value-added French teacher j_{F_A}, and j_{M_B} is coupled with a very low value-added French teacher j_{F_B}, I might wrongly conclude that Math teacher j_{M_B} is relatively better than j_{M_A}.

In the worst case, if Math teachers and French teachers within a school are not randomly allocated, such that better Math teachers and better French teachers are more likely to be allocated to overall better classrooms, the estimated relative value added of the Math teachers would also be biased.

Focus on the corner case: using a link teacher in one subject Importantly, note that these issues could be avoided in one specific sub-case: when the value added of the two French teachers is exactly the same, $\theta_{j_{F_A}} = \theta_{j_{F_B}}$, such that I am only left with the relative value added of the two Math teachers. There is a very small probability that I could consistently compare only Math teachers in classrooms with French teachers of exactly the same quality. However, there is one case in which I can be certain of that: when the French teacher in both classrooms is identical, i.e. $j_{F_A} = j_{F_B}$. In what follows, I propose a method which focuses exclusively on such cases.

Let us for simplicity assume that I can observe a school s with only two classrooms of equal size, $c \in \{c_A, c_B\}$, two Math teachers, $j_M \in \{j_{M_A}, j_{M_B}\}$, and one French teacher j_F. The Math teachers j_{M_A} and j_{M_B} teach respectively in classrooms c_A and c_B, whereas the French teacher j_F teaches in both classrooms. Let us denote the Math teachers j_{M_A} and j_{M_B} as two teachers who are part of a *direct network*, and the French teacher j_F as the *link teacher* of the network - the one through which we can connect the Math teachers to one-another.

Transforming equation 3 by using the fact that teacher value-added is fixed, such that θ_{j_F} is constant across classrooms, I obtain that:

$$\Delta A_{M,c_A,c_B} = (\overline{\mathbf{X}}_{c_A} - \overline{\mathbf{X}}_{c_B}) \cdot (\beta_M - \beta_F) + (\theta_{j_{M_A}} - \theta_{j_{M_B}}) + \eta_{c_A,c_B,M,F} \tag{4}$$

In this case, the condition for unbiasedness of the pairwise difference in teacher value added can be formally written as:

$$\mathbb{E}[(\theta_{c_A,M} - \theta_{c_B,M}) - (\theta_{c_A,F} - \theta_{c_B,F}) | \theta_{j_{M_A}} - \theta_{j_{M_B}}] = 0$$

provided that $\mathbb{E}[(\varepsilon_{c_A,M} - \varepsilon_{c_B,M}) - (\varepsilon_{c_A,F} - \varepsilon_{c_B,F})] = 0$, which is true by assumption.

In this simplified framework with two classrooms, the condition which needs to hold for the estimated value added to be unbiased predictors of the real value added of a teacher is:

Preliminary Condition. The relative Math ability of classroom c_A and c_B should be identical:

$$[(\theta_{c_A,M} - \theta_{c_B,M}) - (\theta_{c_A,F} - \theta_{c_B,F})] = 0$$

While this is a very restrictive condition, it would be useful to provide intuition behind the identifying assumption for a network with many classrooms in a school.

3.2 Direct network observed in multiple classrooms

The previous section concludes with a condition under which one can identify unbiased estimates of teacher value added in a simplified case with one direct network observed in two classrooms.

I now turn to a case where we observe multiple classrooms within a school, which are of equal size. Without loss of generality, let me assume there are four classrooms $c \in \{c_A, c_B, c_C, c_D\}$.

For simplicity, I still focus on a case with a single direct network of Math teachers $j_M \in \{j_{M_{AB}}, j_{M_{CD}}\}$, linked through a single French teacher j_F who teaches in all four classrooms. For the sake of example, let $j_{M_{AB}}$ teach in classrooms c_A and c_B and $j_{M_{CD}}$ teach in classrooms c_C and c_D.

It is easy to show that the Preliminary Condition from the previous section can now be generalised to an assumption necessary for the unbiased identification of the relative value added of two teachers observed together more than once (see Appendix A.1 for details).

Assumption 1. There is no teacher sorting to classrooms within school based on the relative subject-specific ability $(\theta_{c,M} - \theta_{c,F})$**.**

Expressed mathematically,

$$\mathbb{E}\left[\sum_{c \in \{c_0 | j(c_0) = j_{M_{AB}}\}} (\theta_{c,M} - \theta_{c,F}) - \sum_{c' \in \{c'_0 | j(c'_0) = j_{M_{CD}}\}} (\theta_{c',M} - \theta_{c',F})\right] = 0 \tag{5}$$

where $\sum_{c \in \{c_0 | j(c_0) = j_{M_{AB}}\}} (\theta_{c,M} - \theta_{c,F})$ represents the relative Math to French ability of all classrooms of teacher $j_{M_{AB}}$, and $\sum_{c' \in \{c'_0 | j(c'_0) = j_{M_{CD}}\}} (\theta_{c',M} - \theta_{c',F})$ represents the relative Math ability of all classrooms of teacher $j_{M_{CD}}$.

In other words, it should not be the case that the sorting of Math (French) teachers to classrooms is done on the basis of relative Math (French) ability. Note that, importantly, one needs not to worry about sorting on the general level of ability of students.

While this assumption is not directly testable for unobservable factors, it could be tested with the use of observable factors by examining the patterns of student tracking to classrooms and teacher specialisation into teaching certain types of students. As teacher-student sorting is possible only under the condition that both students are sorted non-randomly into classrooms and that teachers are sorted non-randomly to classrooms, one would need for both of these conditions to hold in order to conclude that the identifying assumption is broken.

The tests of the assumption based on observables for the case of the French 9th grade teachers and students is conducted in Appendix B. While ideally, one could test for sorting based on past grades, I do not have access to such data. For this reason, I take advantage of other observable student characteristics which are highly correlated with student grades, such as a student's socio-economic status, scholarship status, age, gender, takeup of advanced classes[8], and region of birth. I predict the student-level difference between her Math and French grades with such observable characteristics. I then test for the two conditions necessary for non-random sorting to occur: tracking of students into classrooms and specialisation of teachers into teaching certain types of tracks.

[8]In particular, I focus on Ancient Greek and Latin which are important for tracking of students in the French system.

To do this, for every school, I regress the predicted difference in grades separately on classroom fixed effects and teacher fixed effects. I compare the average R^2 of such regressions to the distribution of R^2 obtained with bootstrapped data under the null of no sorting based on the difference in Math and French grades predicted by observable characteristics.[9]

I find that the probability of such sorting is lower or equal to the one obtained in the simulated random assignments. In other words, the probability that any type of segregation of students across classrooms or classrooms across teachers has happened by chance is very high. This strongly suggests that I fail to reject the null of no such sorting, which gives supporting evidence to the identifying assumption.

Finally, as shown in Appendix A.1, the analysis requires an additional assumption once we move to a setting where classroom size can vary.

Assumption 2. Each classroom in which a teacher is observed is equally informative of the teacher's value added.

In order to ensure that small classroom size does not add noise to the estimates of value added, I restrict the analysis to only classrooms with more than 15 students. Panel (b) of Figure F.7 confirms that varying the threshold for minimum classroom size does not impact significantly the average within-school standard deviation in estimated value added. In addition, I weigh the effect of observable classroom characteristics $\overline{\mathbf{X}}_c$ by classroom size (details in Section 4).

Finally, in Appendix C I also test that in cases where we observe a network of two teachers multiple times, weighting each observation by the number of students shared by the two teachers does not change estimates of value added. Indeed, the Spearman correlation of teacher value added estimates with and without differential weighting is higher than 0.99 and the average within-school standard deviation in value added is almost unchanged. This confirms that, provided that we exclude very small classrooms which are likely the cause of reporting errors, classrooms may be considered as equally informative of a teacher's value added.

3.3 Direct network observed in multiple classrooms at different times

So far, I have assumed that all classrooms are observed over the same year t. It is useful to explain how complexifying the analysis to a repeated cross section would affect the assumptions of the model.

I now consider a case where I observe four classrooms $c \in \{c_{A,t_1}, c_{B,t_2}, c_{C,t_3}, c_{D,t_4}\}$ observed

[9] Following the same procedure, I also test for possible sorting on common ability, even though such sorting would not violate the identifying assumption. I do indeed find evidence that sorting on common ability is of high probability.

in 4 different periods, $\tau \in [t_1, t_4]$. For simplicity, I still focus on a case with a single direct network of Math teachers $j_M \in \{j_{M_{AB}}, j_{M_{CD}}\}$, linked through a single French teacher j_F who teaches in all four classrooms. For the sake of example, let $j_{M_{AB}}$ teach in classrooms c_{A,t_1} and c_{B,t_2} and $j_{M_{CD}}$ teach in classrooms c_{C,t_3} and c_{D,t_4}.

Taking into account the year of observation, I can rewrite the initial model to:

$$\overline{A}_{c,f,\tau} = \overline{\mathbf{X}}_c \beta_f + \gamma_{c,f} + \theta_{j_f,\tau} + \theta_s + \varepsilon_{c,f,\tau}$$

where the true functional form of teacher effects $\theta_{j_f,\tau} = \theta_{j_f} + \alpha_{j_f,\tau}$, which is time-specific for each period τ due to the presence of the additively separable component $\alpha_{j_f,\tau}$.

I derive the necessary assumption for unbiasedness of the results in Appendix A.2. I show that the results obtained by my proposed methodology would provide me with an estimated pairwise difference of value added that is equal to the true difference given the following condition.

Preliminary Condition. Provided that Assumptions 1, 2 and 3 hold, the estimated relative value added is unbiased if:

$$\mathbb{E}\left[\sum_{t,t'}(\alpha_{j_M,t} - \alpha_{j_{M'},t'}) - \sum_{t,t'}(\alpha_{j_F,t} - \alpha_{j_F,t'})\right] = 0 \tag{6}$$

where $t \in \{t_1, t_2\}$ and $t' \in \{t_3, t_4\}$.

Note that this condition would hold under the strict assumption of a zero mean of individual time components, $\mathbb{E}[\alpha_{j_f,\tau}] = 0$. However, I can provide a more lax assumption if I provide a functional form for $\alpha_{j_f,\tau}$. In particular, I assume that $\alpha_{j_f,\tau}$ is some function of teacher experience, such that $\alpha_{j_M,\tau} = f(exp_{j_M,\tau})$ and $\alpha_{j_F,\tau} = g(exp_{j_F,\tau})$. Provided that this is the case, I can control for this difference in difference in years of experience of a teacher in a flexible manner. This leads me to the next assumption of the paper.

Assumption 3. The time component of value added is additively separable from the intrinsic value added of a teacher and it is a function of years of experience.

Given this assumption, since a teacher's years of experience are observable, one can simply control for $f(exp_{j_M,\tau})$ and $g(exp_{j_F,\tau})$ using fixed effects of experience.

3.4 School network with many teachers within one school

Once I allow for the existence of many classrooms and many teachers within a school, one needs to compare all Math (French) teachers to each other, in order to uncover the entire value added distribution within a school.

Complete school networks At this point it would be useful to introduce some additional vocabulary in order to simplify the discussion.

In the previous section, I have already discussed the meaning of a *direct network* and of a *link teacher*. As a reminder, two teachers from the same subject are said to be in a *direct network* if they can be compared because they teach in the same classrooms as a unique *link teacher* from another subject.

Let me denote an *indirect network* as two teachers from the same subject who do not have a common *link teacher* but can be compared to each other through their connection to teachers in a *direct network*, in other words, by transitivity. Let us further denote a *school network* as the total of all teachers in a subject within a school. Finally, let me make the distinction between *complete school networks* and *incomplete school networks*: I denote a *complete school network* as a *school network* in which all teachers from a subject can be compared to each other, either because of their *direct networks* or *indirect networks*.

To better grasp the idea behind these terms, it is worthwhile to examine Figures 1 and 2. Both figures represent hypothetical *school networks* of Math and French teachers, represented in blue and yellow nodes, respectively. An edge connecting a Math and a French teacher represents that the two teachers are observed together in a classroom.

Figure 1 provides an example of a *complete school network* which consists of 6 Math teachers (M_1 to M_6) and 6 French teachers (denoted F_1 to F_6). To see this, note that teachers M_1 and M_2 are in a *direct network*: they can be directly compared to each other as they are both observed in classrooms with French teacher F_1. Furthermore, Math teachers M_2 and M_3 are also in a direct network *direct networks*: through their observation with *link teacher* F_2. Note further that Math teachers M_1 and M_3 are not in a *direct network*, but they are however in an *indirect network* as they can be compared by transitivity, as they are both in a *direct network* with M_2. Similarly, all Math teachers can be compared to each other either through direct or indirect networks. Therefore, one can uncover the entire distribution of Math teacher value-added in this school. The logic extends similarly to French teachers.

By contrast, Figure 2 provides an example of an *incomplete school network*. Math teachers M_1, M_5 and M_6 can be compared to each other, but they cannot be compared to M_2, M_3 and M_4 as there is no one teacher that is in a *direct network* with another teacher from both of these subgroups. Therefore, one cannot construct the full distribution of Math teacher value-added in this school. The logic follows for French teachers as well.

It follows that in order to estimate teacher value-added one needs to have a *complete school network*.

While in theory one would prefer to have a *complete school network* of teachers every year (which would be equivalent to having year-specific teacher effects), in practice this might be complicated due to a restricted school size. In particular, if a school has very few classrooms per year, it follows that there would be few teachers per year, which entails few networks. At best, this would lead to very noisy coefficients of value-added, as the pairwise comparison between two teachers would be computed off one or few classrooms. At worst, the issue would lead to *incomplete school networks* and thus an impossibility to measure value added.

For this reason, Assumption 3 plays a very important role, as it allows to construct networks based on a longer period than one year, and in particular use the entire available period of networks of a teacher to identify her value added, provided that one controls for experience in a flexible manner.

Thus, no further assumptions are necessary to uncover the entire school distribution of teacher value added in a subject.

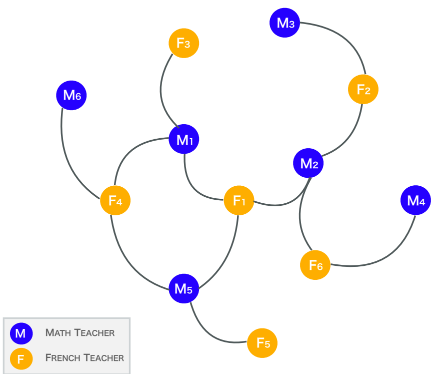


Figure 1: Complete school network

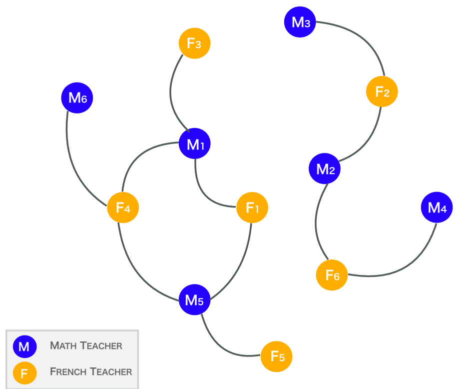


Figure 2: Fragmented school network

3.5 System of networks with many teachers within many schools

The method allows to measure value-added of teachers only within school. What this entails is that one cannot directly compare the coefficients of two teachers from two different schools, or of the same teacher observed in two different schools.

This is important as the method does not require any additional assumptions on the distribution of teachers across schools, which is much less likely to be random than the within-school distribution of teachers across classrooms.

In fact, the only additional assumption required by extending the model to multiple schools has to do with school representativeness. It is only necessary if one wants to say something about the average variability in teacher effectiveness.

Assumption 4. The schools which exhibit a complete school network are representative of the school system.

The assumption is once again easily testable for observable school characteristics. For the case of the French system of schools, I provide details in Section 6.3 and Table E.8 which show that this is the case for the treated French schools (86% of all public middle schools) based on a large number of school-level average student and teacher characteristics. In particular, while there appear to be some statistically significant differences between treated schools and the average school, these are economically insignificant.

4 Empirical strategy

The empirical strategy follows very closely the model discussed in the previous section.

First stage I begin by estimating for each student i and each subject $f \in \{M, F\}$ a regression of the sort:

$$A_{i,f,j_f,t} = \mathbf{X}_i\beta_f + \alpha_{j_f,t} + \theta_s + \varepsilon_{i,f,j_f,t}$$

where $\mathbf{X}_i$ is a vector of observable student characteristics, $\alpha_{j_f,t}$ are fixed effects for years of teacher experience, and θ_s are school fixed effects. This is preferred to directly averaging on a classroom level as it allows to give more weight to larger classrooms in estimating the coefficients. The estimation requires the inclusion of the school fixed effects θ_s in order to follow the identification based on within-school comparisons of teachers.

Second stage I take the student level difference of the residuals $(\widehat{\varepsilon}_{i,M,j_M,t} - \widehat{\varepsilon}_{i,F,j_F,t})$ and aggregate it to a classroom level denoted $(\widehat{\varepsilon}_{c,j_M} - \widehat{\varepsilon}_{c,j_F})$. In order to compute for each observation n of a direct network of any two Math teachers j_{M_A} and j_{M_B} and any given French link teacher j_F, I take the difference:

$$(\widehat{\varepsilon}_{c_A,j_{M_A}} - \widehat{\varepsilon}_{c_A,j_F}) - (\widehat{\varepsilon}_{c_B,j_{M_B}} - \widehat{\varepsilon}_{c_B,j_F}) \equiv \widehat{\varepsilon}^{n}_{j_{M_A},j_{M_B}}$$

I then average across all observations of a single direct network of two teachers in a subject, such that the relative value-added of the Math teachers j_{M_A} and j_{M_B} is computed as:

$$\widehat{\theta}_{j_{M_A}} - \widehat{\theta}_{j_{M_B}} = \frac{1}{N}\sum_{1}^{N} \widehat{\varepsilon}^{n}_{j_{M_A},j_{M_B}}$$

and similarly for a direct network of any two French teachers linked by any given Math teacher.

Moving from pairwise comparisons to a distribution Finally, as I obtain only pairwise comparisons between teachers in a direct network, it is necessary to discuss how one can use these comparisons to also compare teachers in an indirect network, and furthermore turn these comparisons into single coefficients for each teacher which would allow for comparing teachers along the entire distribution of value added within a school. Problems with this arise as one undoubtedly is faced with the issue of an overdetermined system of these linear equations, and thus the possibility to find many different estimates for value added of a teacher based on the algorithm used to solve this system.

I propose to solve this system by OLS. To make this clearer, imagine a system of pairwise comparisons of three Math teachers, such that $\widehat{\theta}_{jM_A} - \widehat{\theta}_{jM_B} = a$, $\widehat{\theta}_{jM_C} - \widehat{\theta}_{jM_B} = b$, and $\widehat{\theta}_{jM_A} - \widehat{\theta}_{jM_C} = c$, but $c \neq a - b$. Note that I can write out this system in a matrix form $Ax = v$, such that I obtain:

$$\begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} x = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Thus, instead of trying to equate each equation to zero $Ax - v = 0$, the use of OLS minimises the sum of squared distances from zero and allows us to obtain a single estimate for each teacher's value added.

Note that the method requires choosing arbitrarily a teacher to whom to allocate a value-added of 0, i.e. by dropping a column from the matrix **A**. I choose to arbitrarily allocate a value-added of 0 to the first teacher in a subject observed in my school-level dataset. Naturally, an implication of this is that by construction, the mean value-added within a school does not have any meaning, contrary to the results from the typical estimation of within-school value added where the zero-mean represents the believed true mean of value added within a school. Thus, I can de-mean the value-added estimates within-school, such that all school means are shifted to zero, making deviations from the mean comparable.

I apply these methods within school and calculate the standard deviation of value-added measured within each school. I then take the average within-school variability in teacher value-added in order to provide an idea of the average possible gains in student grades from moving a student to a better teacher within the same school.

5 Simulations

Before applying the method empirically on French data, I conduct Monte Carlo simulations to illustrate the performance of the method in terms of recovering the true value added effects under three possible scenarios of teacher-student sorting within school (discussed in Section 3). First, I begin from a simulated network of teachers who are sorted to students randomly. Second, I simulate a type of sorting where students are sorted into a classroom and to a Math teacher based on their general level of ability (i.e. the common component of ability between Math and French). Third, I simulate a type of sorting where students are sorted into a classroom and to their Math teacher based on their Math-specific ability. For comparison, I provide the results of these simulations for the traditional method as well.[10]

To better compare the predictive power of the two methods, I focus on three measures. First, in order to find how accurate is the estimated rank of each teacher in the value added distribution resulting from my method, relative to that of the traditional method, I compute the Spearman correlation between each of the teacher value estimates and true value added. Second, in order to find how precise are my value added estimates compared to the traditional estimates, I compute the Mean Squared Error (MSE) of the estimates, compared to the true value added. Third, in order to assess the presence of bias in addition to precision, I provide graphical evidence of the estimates of value added from each model and each sorting scenario, each averaged over 1,000 Monte Carlo simulations, compared to the true parameters.

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, number of students, and number of teachers, over a period of 8 years (see Table 1). 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.

Characteristic	**Size**
Number of classrooms	25
Average classroom size	24
Number of Math teachers	5
Number of French teachers	6

Table 1: Simulated school

Note: The table depicts the characteristics of the average school in my sample of schools for a total period of 8 years. These characteristics are used for the purpose of simulating the average school.

I use 1,000 Monte Carlo replications per scenario per method, where I keep the distribution of true value added fixed but I reassign students to classes and classes to teachers, and redraw all other random variables.

5.1 Setup

To focus more explicitly on the problem of model predictions, I restrict the DGPs to a narrow set of relatively idealised conditions. I assume that test scores of a student are only a function of overall ability, subject-specific ability and teacher value added - in other words there is no other idiosyncratic factor, or other observable characteristics such as socio-economic status,

[10]See details of the precise estimations used in Appendix D.4

which affect grades. I assume further that there are no time effects, no peer effects, and teacher value added is constant.

The random variables used in the simulations of my method are therefore precisely the teacher value added, θ_{j_f}, overall student ability, θ_i, and subject-specific ability $\theta_{i,f}$. In addition, for the simulations of the traditional method, I generate a lagged test score variable $A_{A_{i,f,j_f,t-1}}$. All four variables are drawn from Normal distributions, such that lagged test scores are by construction correlated with the sum of ability $(\theta_i + \theta_{i,f})$.

Simulated scenario 1. Random sorting First, I simulate random sorting between teachers and students. To simulate student grades in Math, A^s_{i,M,j_M}, I draw "true" teacher value added $\theta_{j_M} \sim \mathcal{N}(0, 0.134^2)$, overall ability $\theta_i \sim \mathcal{N}(0, 0.339^2)$ and Math-specific ability $\theta_{i,M} \sim \mathcal{N}(0, 0.339^2)$. The choice of standard deviations follows the findings of Chetty et al. (2014a). More specifically, I take the standard deviation of teacher value added found in Chetty et al. (2014a) and its share of the total variance - 0.07. I then assume the rest of the total variance is equally shared between overall ability and Math-specific ability (i.e. with shares 0.47 and 0.47). In Appendix D.2, I show that the results are robust to varying these shares of the variance. Specifically, I first vary the respective shares of overall ability and Math-specific ability, keeping the weight of teacher value added (respectively 0.23, 0.69, 0.07 and 0.69, 0.23, 0.07), and show that the higher the share of the common ability is in the total variance, compared to the share of Math-specific ability, the better the MSE of my method is. In addition, following Guarino, Reckase, and Wooldridge (2015), I show that increasing teacher value added's share of the variance to 0.21 yields similar results.

Using these random generated variables, I compute the Math grade[11] of a student i as:

$$A^s_{i,M,j_M} = \theta_{j_M} + \theta_i + \theta_{i,M} \tag{7}$$

The DGP for the French grade of a student A^s_{i,F,j_F} is the equivalent with one exception. I choose to vary the French-specific ability's standard deviation, $\sigma_{\theta_{i,F}}$, in order to vary the correlation between A^s_{i,M,j_M} and A^s_{i,F,j_F}. In other words, I exploit the fact that the French grade is used for the Math teachers' value added derivation in my method, but it is not used in the traditional method, in order to vary the predictive power of my method for a given predictive power of the traditional method. This allows me to compare the predictive power of both methods for different levels of correlation between the residualised student grades in two subjects, which is important for external validity. I also focus more specifically on the correlation $\rho_{A^s_M, A^s_F} = 0.58$, which is equivalent to the correlation between the Math and

[11] One can think of it as the residualised student grade (net of observable student characteristics).

French grades net of observable characteristics in my data.[12]

The final random variable that I generate is the lagged exam score of a student in Math, $A_{i,M,j_{M'},t-1} \sim \mathcal{N}(0, \sigma^2_{(\theta_i+\theta_{i,M})})$, which is used in the traditional method. $A_{i,M,j_{M'},t-1}$ is correlated with the total ability of a student in Math, $(\theta_i + \theta_{i,f})$, by a factor $\rho \in [0.1, 0.9]$. As the lagged grade is not used in my method, I exploit this range of ρ to vary the predictive power of the traditional method for a given predictive power of my method. This allows me to compare the predictive power of both methods for different levels of correlation between lagged student grade and total ability. I focus more specifically on the correlation $\rho = 0.8$, which is on the upper bound of the reported correlation in the literature.[13]

Simulated scenario 2. Sorting on common ability Second, I simulate a setting where students are sorted to Math teachers based on their common ability θ_i. In particular, I use an extreme example where θ_i is no longer generated from the DGP specified in the random sorting scenario, but it is instead generated in such a way that it is 1-to-1 correlated with θ_{j_M}, through an equation $\theta_i = a \times \theta_{j_M} + b, a > 0$.[14] This would mean that better Math teachers are sorted into classrooms with better students.

As shown in Section D.4, θ_i cancels out in my method. However, in the traditional model, the use of the past grade does not fully control for the non-random sorting as the correlation between $A_{i,M,j_{M'},t-1}$ and $(\theta_i + \theta_{i,M})$, ρ, is lower than 1.

Simulated scenario 3. Sorting on Math-specific ability Third, I simulate a setting in which students are sorted to Math teachers based on their Math-specific ability $\theta_{i,M}$. More specifically, I use an extreme example where $\theta_{i,M}$ is no longer generated from the DGP specified in the random sorting scenario, but it is instead generated in such a way that it is 1-to-1 correlated with θ_{j_M}, through an equation $\theta_{i,M} = a \times \theta_{j_M} + b, a > 0$.[15] This would mean that better Math teachers are sorted into classrooms with students who are better specifically in Math.

It can be shown that in this scenario, each pairwise comparison of teachers in a network

[12] Note that the standard deviation σ_{θ_i} is still derived depending on the Math teacher's value added and the Math-specific ability, as outlined above. This is not important for the simulation of my method, as θ_i would cancel out in either way. It is also not important for the traditional method, as A^s_{i,F,j_F} is not used for the derivation of the Math teacher's value added. Finally, $\theta_{j_F} \sim \mathcal{N}(0, 0.0.098)$, as in Chetty et al. (2014a).

[13] Rothstein (2009) shows that the correlation for actual reading grades between classes is between 0.7 and 0.8.

[14] Note that such a DGP would lead to a mean $\mu_{\theta_i} = a\mu_{\theta_{j_M}} = 0$ and a variance $\sigma^2_{\theta_i} = a^2\sigma^2_{\theta_{j_M}}$. For comparability with the previous scenario, I thus rescale the new variable θ_i in order to obtain the same $\sigma^2_{\theta_i}$ as in the case of no sorting.

[15] Note that such a DGP would lead to a mean $\mu_{\theta_{i,M}} = a\mu_{\theta_{j_M}} = 0$ and a variance $\sigma^2_{\theta_{i,M}} = a^2\sigma^2_{\theta_{j_M}}$. For comparability with the previous scenario, I thus rescale the new variable θ_i in order to obtain the same $\sigma^2_{\theta_i}$ as in the case of no sorting.

would be biased by a factor $(1+a)$ (see Appendix D.4 for details). As in Simulated scenario 2, the traditional model also suffers from bias.

5.2 Simulation results

Table D.4 and Figures D.2, D.3 and D.6 in the Appendix depict the results for the MSE and Spearman correlation between each of the models' estimates and the true value added parameters for the three cases: no sorting, sorting on common ability and sorting on subject-specific ability. Table D.4 focuses on the specific case with correlation between past grade and student ability of 0.8, and correlation between residualised Math and French grade of 0.58 (i.e. the empirically identified statistics based on Rothstein (2009) and the data on 9th grade students in France, respectively). Figures D.2, D.3 and D.6 vary the two types of correlation to provide a distribution of results. In addition, panels (c) and (d) of each figure provide a better idea of the individual value added estimates and allow to discuss potential bias.

Precision of estimates As seen in Table D.4, in the case of no sorting, the MSE of both models is very small and close to each other - such that the square root of the MSE of my method's estimates is 0.005 and that of the traditional model's estimates is 0.003). In Figure D.2, one can see that the difference between the MSE of my method and traditional method is lower the higher the correlation between Math and French grades is. Varying the level of correlation between student ability and past Math grade does not change the difference substantially (as seen by the small difference across lines).

Once (extreme) sorting on common ability is introduced, the MSE of my method stays unchanged, consistent with the theoretical prediction of the method, the MSE produced from my method stays unchanged, while the MSE of the traditional method grows substantially (its square root is 0.244). As seen in Figure D.3, my method outperforms the traditional in terms of MSE consistently, irrespective of the correlation between Math and French grades, and decreasing with the correlation between student ability and past Math grade.

When I move to (extreme) sorting on subject-specific ability, as expected the MSE of the my method grows substantially (the square root is 0.464), while the square root of the MSE of the traditional method is similar to the one obtained given sorting on common ability - 0.246. It is visible from Figure D.6 that the traditional method indeed consistently outperforms my method in terms of MSE. The difference between the two MSEs is slightly decreasing with the increase in the correlation between Math and French grades and is increasing with the correlation between student ability and past Math grade.

Precision of the ranking of estimates Irrespective of the presence of sorting, the prediction of the rank of each teacher's value added in the distribution of value added is equally good for both my method and the traditional method, and in fact is perfect for the benchmark values of correlation chosen. As visible from Figures D.2, D.3 and D.6, the difference between the Spearman correlations is always zero, irrespective of the correlation between ability and past grade and the correlation between Math and French grades.

	NE	**BE**	**Δ (NE-BE)**
No sorting			
Sqr. Root of MSE	0.005	0.003	0.002
Spearman corr.	1.00	1.00	0.00
Sorting on common ability			
Sqr. Root of MSE	0.005	0.244	-0.239
Spearman corr.	1.00	1.00	0.00
Sorting on subject-specific ability			
Sqr. Root of MSE	0.464	0.246	0.218
Spearman corr.	1.00	1.00	0.00

Table 2: Model performance comparison

Note: This table depicts the square root of the mean squared error (MSE) of the Traditional model (BE) compared to the true value added parameters and of the Network model (NE) and the true value added parameters. It also shows the Spearman correlations between the BE model and the true value added parameters, and the NE model and the true value added parameters. The referenced BE model is estimated for a correlation $\rho = 0.8$ between past student grade and student ability $(\theta_i + \theta_{i,f})$. The referenced NE model is estimated for a correlation 0.58 between A^s_{i,M,j_M} and A^s_{i,F,j_F}. The case of *No sorting* does not impose any correlation between value added and student ability. The case of *Sorting on common ability* imposes a correlation of 1 between the value added of the Math teacher and the common ability factor θ_i, specifically by imposing a structure $\theta_i = a \times \theta_{j_M} + b$. Finally, the case of *Sorting on subject-specific ability* imposes a correlation of 1 between the value added of the Math teacher and the Math-specific ability factor $\theta_{i,M}$, specifically by imposing a structure $\theta_{i,M} = a \times \theta_{j_M} + b$.

Bias As visible from panels (c) and (d) of Figure D.2, in the absence of sorting there is no observable bias in the estimated value added. In panel (c), one can notice that, given variation in the correlation between Math and French grades, while the precision of my method's estimates changes, there is no sign of bias in one direction or another.

As shown in Figure D.3, it is worth noting that once I introduce sorting on common ability, while there is no bias in my method's estimates, there is positive bias in the traditional method's estimates. In other words, the value added of better teachers (those above the reference teacher with zero value added) is overestimated, while the value added of worse teachers is underestimated, compared to the true values. In panel (d), one can see that the level of bias is higher the lower the correlation between student ability and past student grades is. This is consistent with the theoretical prediction outlined above.

Finally, once I introduce sorting based on the subject-specific component of ability, one can note in Figure D.6 that indeed both my method and traditional method produce estimates which are biased upwards. However, the bias is lower in the traditional method, specifically for high levels of correlation between ability and past grades.

All of these findings indeed confirm that my method performs well compared to the traditional in the case of random sorting of students to teachers. If sorting is based on common ability, my method significantly outperforms the traditional method. Finally, as confirmed in Appendix D.2, in all three scenarios, the larger the share of common ability is in the total variance of student grades, compared to the share of Math-specific ability, the lower the MSE of my method is.

6 Institutional setting and data

As this paper uses French data in order to apply the methodology empirically, this section briefly discusses the French institutional setting and provides details on the data used.

6.1 Institutional setting

Due to the use of French data for the test of the proposed value-added methodology, it is worthwhile to have an overview of the French institutional setting. In this section, I briefly discuss the educational system, focusing particularly on middle schools in France and the specificities of the teaching career, as these would be important for better understanding the distribution of teachers in France.

6.1.1 Overview of the French educational system

The educational system is split in public and private schools, such that about 80% of secondary school students attend public school.[16] Due to the prevalent presence of public schools, the wider differences in teacher remuneration, recruitment and advancement, as well as the higher concentration of high socio-economic status students in private schools, I focus exclusively on public schools in the empirical tests of my identification strategy.

Education in France is compulsory from the age of 6 until the age of 16, when students finish middle school - generally, following the obtainment of National Brevet Diploma (DNB).[17] After middle school, students have a choice to enter general, vocational or technical training.

[16] Private schools are either to an extent publicly-funded (mostly catholic schools; they follow the curriculum of the Ministry of Education at the exchange of teachers being paid by the State) or privately funded (they do not follow the official curriculum and depend on strong financial participation of families).

[17] There are two, less common, alternatives to the DNB: the General Training Certificate (CFG) and the IT and Internet Certificate (B2i-IT). The former is intended for adapted pupils (SEGPA) and those in regional adapted education institutions (EREA). The latter is an attestation of skill level acquired by the pupil in mastering multimedia tools and the Internet.

About half of middle-school graduates continue with the general track, which offers the most opportunities after graduation and is a necessary step for students who wish to continue with higher education.

The French educational system does not rely on annual standardised testing. It does, however, rely on standardised testing in specific years of the program. One such important standardised exam occurs at the end of the 9th grade - the final middle school year. In order to obtain the DNB at the end of middle school, in addition to continuous assessment, which weighs for 400 out of the 700 points, all students are required to sit a set of national standardised exams at the end of 9th grade which weigh for the remaining 300 points. These exams (hereafter *DNB*) test the knowledge, skills and culture acquired throughout middle school in Math, French language and other subjects. While pursuit in high school is not conditional on the obtained DNB grade, the exams are nevertheless considered of high importance for the smooth passing of the 9th grade.[18] As teachers prepare their students for the exams over the entire 9th grade, the exam grades should indeed reflect well the effectiveness of teaching.

6.1.2 The teaching career

Generally speaking, there are two types of teachers - teachers on a tenure track and contractual teachers.

Individuals wishing to become tenured-track teachers undergo a year of intensive pedagogical preparation at university for a competitive national exam. There are different certification exams depending on whether one wishes to teach in nursery and primary schools (first degree) or in middle schools and high schools (second degree). The most common certification for those who wish to teach in second degree is the Certificate of Aptitude for Teaching Secondary Education (CAPES). A more advanced certification alternative - the Agrégation, allows to teach in high schools and higher education and comes with additional perks such as higher salary and fewer teaching hours. However, it is less common among teachers in middle school due to the difficulty and selectivity of the certification exams. In both types of certifications, successful candidates become trainee teachers for one to two years, and are allocated to a school and an experienced teacher within that school who acts as a tutor. Once the internship is completed, the teacher is officially appointed as a tenured teacher[19] and becomes a civil servant.

An alternative way to teach in both first and second degree is to become a contractual

[18] The diploma is also required for certain public service competitions and professional internships. Furthermore, scholarship holders on social criteria who obtain a "Good" or "Very Good" grade can apply for a merit scholarship during their high school career. Its importance seems tangible for students, who have the option to decide to repeat the 9th grade in order to retake the exam, in case they have missed it the first time.

[19] In the case of teachers with Agrégation, candidates need to also go before a jury to validate their diploma.

teacher. According to data from the Statistical office of the French Ministry of Education, the DEPP, about 8% of all teachers in secondary education are contractual teachers. Contractual teachers have not passed the recruitment competitions organised by the Ministry of Education and are simply required to have studied for at least 3 years after completing their general secondary education. Generally, their mission is to replace a tenured teacher who is on leave, or to fill a vacant position in a public school which cannot be filled by a tenured teacher. This could explain why in my data, a contractual teacher has a 19% higher probability of teaching at a disadvantaged school.

Due to the nature of their contract, they are bound to stay at a school for a maximum period of 3 years. Indeed, in the sample of 9th grade Math and French teachers, the average time for which a contractual teacher stays in a school is 1.8 years, compared to 6.5 years for the average tenured teacher. The salary of contract teachers is also tied to a different pay scale and is generally lower than the salary of tenured teachers.

As the period of observation within a school of contractual teachers is generally lower, and as one might argue this could affect such teachers' motivation and incentives on the job quite differently than those of a tenured teacher, I do not focus on contractual teachers in the empirical exercise in order to lower the potential noise coming from their value added.

6.2 Data sources

This study relies on administrative data from the statistical office of the French Ministry of Education (DEPP) for the period between 2009 and 2018.[20] Specifically, to construct my final dataset, I utilise two main databases: *RELAIS* - a database with rich information on teachers, and *FAERE* - a database on students.I focus exclusively on public schools in the 26 *académies* in Metropolitan France, concentrating on students in the 9th grade and their teachers in Math and French.

The database *RELAIS* contains a large amount of individual teacher-specific information, such as gender, date of birth, qualification, teaching status, subject taught, rank on the wage scale and pedagogical grades. Each teacher is also associated with the school(s) in which she teaches during each year of her career.[21] Importantly, for every year of their career, the *RELAIS* database provides a classroom identifier within the school in which she teaches, indicating the specific classroom in which they teach.

The database *FAERE* contains individual data on students, such as gender, date of birth, city of birth, current residence, occupation and working status of parents, scholarship status,

[20]Note that currently, the study does not use information from the 2016-2017 school year as *DNB* grades are not yet available for use.

[21]All of this information is available in the files *MEN*, *EMF* and *AIR*, which can be merged with the use of unique teacher identifiers.

choices of optional courses, grades in the national standardised examination *DNB*, as well as grades of the continuous assessment during the 9th grade, and importantly, the classroom in which the student is taught.[22] The latter allows to match each teacher to their respective students. I standardise the *DNB* grades by year and subject to allow for comparability across years. In addition, I use parent profession in order to compute each student's socio-economic status (SES) based on guidelines provided by the Ministry of Education, split in four categories: very high SES (*Très favorisée*), high SES (*Favorisée*), medium SES (*Moyenne*) or low SES (*Défavorisée*). Specifically, I classify each student according to the occupation of their parent with the higher rank.[23]

In order to reduce the potential noise in estimates, I exclude very specific classroom types, such as classrooms for students with mental disabilities or integration classrooms for non-French speakers. I furthermore exclude classrooms with very few students (less than 15) and schools which are not observed for more than 2 years, both of which are likely reporting errors. I also exclude schools with very few classrooms (less than 5 in total over the years), as those have too few teacher networks to provide precise estimates. Finally, I choose to focus only on tenured (as opposed to contractual) teachers in order to make sure we have a sufficient number of teacher observations per teacher within a school.

6.3 Descriptive statistics

Appendix E provides descriptive statistics of the variables used in this paper for the sample that enters the analysis, split into characteristics of 9th grade students and characteristics of Math and French teachers in 9th grade.

As seen in Table E.7, 19% of students in 9th grade are placed in a school labelled as disadvantaged, and 23% are scholarship recipients. The pass rate of 9th grade stands at 76%. About 97% of students are French nationals. Finally, according to the chosen classification of socio-economic status, 54% of students recorded as having a medium (33%) or low (21%) SES, compared to 19% of students having a high SES and 27% of students - a very high SES.

Math and French teachers' characteristics are shown in Tables E.5 and E.6, respectively. Math (French) teachers in the sample are on average 40 (41) years old, with about 11 (12) years of experience. Only 9% (8%) of Math (French) teachers teaching in 9th grade have an advanced certification level (*Agrégation*) which is expected since, as discussed above, the

[22] The information exists in the *Apprenant* and *DNB* databases and can be merged through unique student identifiers.

[23] As labour decisions of each parent may depend on the labour decisions of the other parent, a household's socio-economic status is likely best represented by the higher earner's socio-economic status. I thus choose to use this classification rather than the classification used in the Ministry, which relies exclusively on the father's rank, as the higher earner may be the mother.

majority of *Agrégés* typically continue to teach at high school or higher education. The average gross monthly salary base for Math (French) teachers is about 3,000 euros (3,020).[24] The sample is almost evenly split between Math and French teachers, with about 48% of teachers teaching Math. Finally, 21% of Math teachers and 21% of French teachers teach at a disadvantaged school.

For the total of middle schools in Metropolitan France in our sample, an incomplete school network is observed in only 14% of schools. We can compare the observable characteristics of the treated schools to the average characteristics in the school system in order to make sure that the treated schools are representative. As seen in Table E.8, while there are some statistically significant differences that define the treated schools, but the economic significance of such differences is very insignificant.

Beyond the issue of school network completeness, it is also important to study the characteristics of the teacher networks in schools that have a complete network. As we can see in Table E.9, we observe about 40,000 unique teacher networks in Math and about 60,000 in French. The average Math (French) teacher is part of 4.4 (5.4) networks within a school, and each network is observed on average 5.6 (4.9) times, which corresponds to 135 (117) students per couple.

7 Results

The coefficients of the first stage are shown in Table F.10. After obtaining the estimates of the pairwise differences in value added of pairs of teachers, I apply the matrix method for deriving single teacher effects, as described in Section 4.

Using this method, I find that for a 1 SD increase in Math (French) teacher value-added within a school, student scores benefit on average by 0.174 SD (0.163 SD).[25]The effect is substantial: what it means is that if the average student in a school moves from a Math (French) teacher in the 5th percentile of the school's value-added distribution to a teacher in the 95th percentile, his Math (French) grade would improve by 0.58 SD (0.54 SD). Given the distribution of grades in each subject, this is equivalent to a 14 percentage points increase in the student's Math grade and a 10 percentage points increase in his French grade.

The coefficients are placed on the higher bound of the estimates found in the existing literature on US middle schools, which typically lie between 0.10-0.15 SD for Math and 0.05 and 0.15 SD for literature (Jackson, 2014,Bacher-Hicks & Koedel, 2022), with higher estimates

[24]Note that I compile this information based on the public pay scheme for teachers published on the website of the Ministry of Education.

[25]Note that the Pearson and Spearman correlations between the derived value added estimates from this empirical strategy and an alternative strategy where one does not control for experience fixed effects is above 0.98. This indeed confirms that the importance of the drift in value added is very small.

for Math than for literature (Lefgren & Sims, 2012; Condie et al., 2014). In fact, they are closer to estimates found in developing countries (e.g. Bau & Das, 2020; Buhl-Wiggers et al., 2017 find coefficients between 0.11 and 0.19 SD).

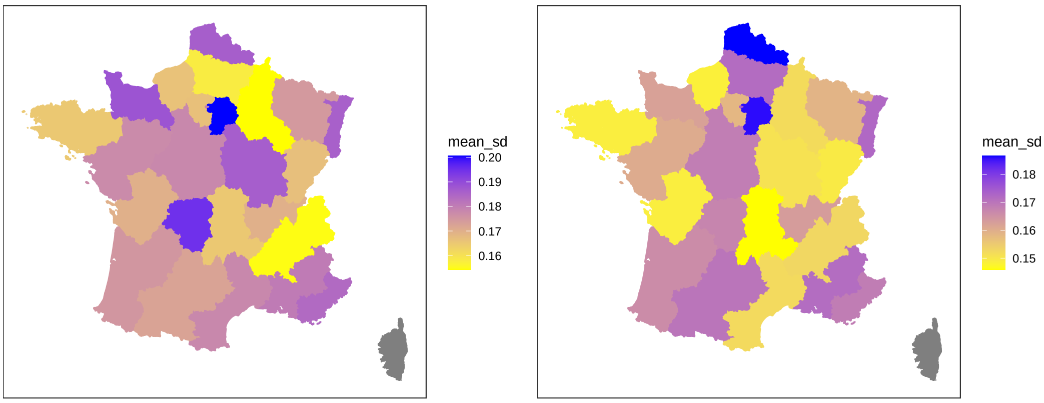
These results are robust to changes in the restrictions on the minimum number of classrooms within a school and the minimum classroom size (see Figure F.7 in the Appendix). I also re-estimate the model weighting for total number of students shared between two teachers in a network and obtain almost identical results (Appendix C).

Heterogeneity There is, albeit small, variation in the within-school variability of value-added estimates by French educational region (see Figures 3). In particular, the highest variation in within-school teacher effectiveness in both Math and French is measured in Creteil, near Paris, the French region with the highest difficulties in attracting teachers and the region with the highest share of newly tenured teachers.[26,27]

Figure 3: Within-school s.d. in value added by region

Note: The figure shows the académie-level average standard deviation in teacher value-added within school. Corsica is excluded from the figure.

(a) Math VA (b) French VA



This is also consistent with the finding that the average within-school variability in teacher value added is higher in more disadvantaged schools. In Figure 4, schools are split into deciles of disadvantaged level, based on the share of low-SES students the school welcomes on average. As seen, while the average variability seems relatively stable in general, it jumps sharply for the last decile of schools, in other words - for the most disadvantaged schools.

[26] According to information from the Pommiers and Lecherbonnier (2022) Le Monde article “Education : comment le système de mutation des enseignants s'est grippé”.

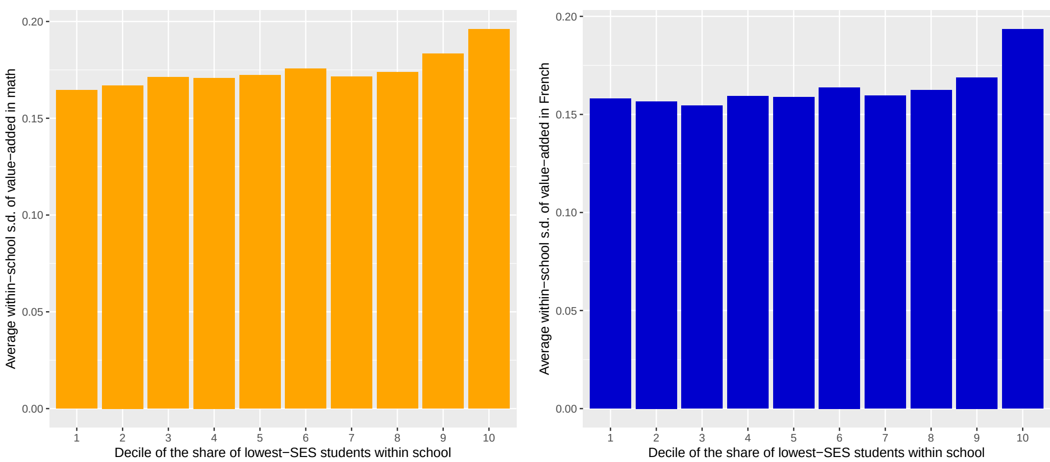
[27] Similar conclusions can be drawn for Lille - the other region with the highest variation in value-added for French teachers.

This finding hints that the differences in teacher quality are particularly drastic in schools which predominantly teach disadvantaged students.

Correlation of value added estimates with observable characteristics In addition to using the value added estimates to provide some average idea of variability in teacher effects, it is interesting to observe if there are certain characteristics which are correlated with these value added estimates. Figure 5 shows the correlations between teacher characteristics and the value added estimates.[28]

Figure 4: Standard deviation of value added by school disadvantaged level
Note: The figure shows the average within-school standard deviation in Math teacher value added by decile of "school disadvantageness level". School disadvantagedness level is defined by computing the share of students with "Low SES" out of all students within the school, and then splitting schools into deciles of this measure, such that the lowest decile represents the schools with the lowest share of low SES students, and vice versa for the highest decile.

(a) Math VA (b) French VA



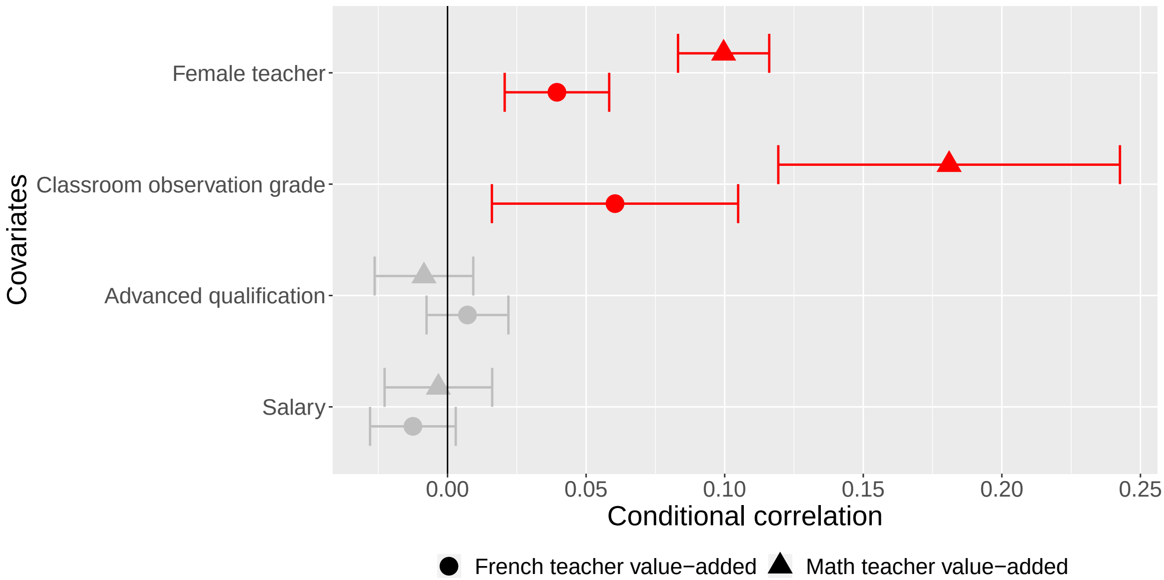
Female teachers have on average higher value added, similar to what is found in the literature (e.g. Aaronson, Barrow, and Sander (2007)). This finding is to a large extent driven by female teachers being better at teaching female students, but the effect persists even when accounting for this relationship.

The classroom observation grade given to a teacher for their pedagogical skills by inspectors sitting in their class is strongly positively associated with value added - in fact, both for Math and French teachers, it is the characteristic that best correlates with the value added estimates. The coefficient of correlation is also in line with that found in the literature (e.g. Rockoff, Jacob, Kane, & Staiger, 2011; Kane & Staiger, 2012; Harris & Sass, 2014; Jacob & Lefgren, 2008).

[28]Note that I repeat the exercise but add all observable characteristics to a regression for a small subsample of teachers for whom there is no missing data in any of the characteristics. Appendix Table F.3 broadly confirms these predictions.

Finally, having an advanced qualification (*Agrégation*) is not associated with having higher value added. Similarly, having a higher salary (which is in itself a function of a fixed set of characteristics such as seniority) is not associated with higher value added either. This is consistent with the idea that remuneration based solely on observable teacher characteristics is independent of considerations of value added.

Figure 5: Correlation between value added estimates and teacher characteristics
Note: The figure shows the conditional correlation between teacher value added estimates and teacher characteristics. The correlations are conditional on school fixed effects to reflect the fact teacher value added estimates are comparable only within school. The characteristics are either fixed or averaged over the entire period in which the teacher is observed within school.



8 Conclusion

This paper propose an alternative method for teacher value added estimation which does not rely on lagged test scores - a measure typically used to control for potential non-random sorting between teachers and students, which might bias value added estimates. Instead, I derive a measure of within-school value added with the use of within-student cross-sectional variation in standardised exam grades in different subjects. Using direct networks of teachers - teachers in the same subject who have been observed in a classroom with the same teacher from another subject - I show how one can compare the relative value added of teachers in a pairwise manner, and how one can uncover the entire value added distribution within a school by transitivity.

I show that certain assumptions need to hold in order to provide unbiased estimates of teacher value added. Firstly, I assume there is no teacher sorting to classrooms within school based on subject-specific student ability. I provide a simple test of the assumption using observable characteristics and indeed show that in my observed setting, this assumption is not a cause for concern. Secondly, I assume each classroom in which a teacher is observed is equally informative of the teacher's value added. I perform several tests which point in that direction, and restrict the sample to large enough classrooms to minimise the potential noise of smaller classrooms. Thirdly, I assume that the time component of value added is additively separable from the intrinsic value added of a teacher and is a function of years of experience. I then take care of this empirically by adding fixed effects for the the years of experience of a teacher in each classroom. Finally, unrelated to unbiasedness, I assume that the schools which exhibit a complete school network (i.e. the schools for which we can conduct the analysis) are representative of the school system. This is necessary to ensure the average result has merit on the aggregate. I conduct balance checks which confirm that while there are some statistically significant differences between the treated schools and the average school, these differences are very small.

To compare the estimator to the one commonly used in the literature, I simulate a representative school and go through three possible types of sorting of students to teachers - random sorting, sorting based on common ability, and sorting based on subject-specific ability. I show that my method performs slightly worse but generally quite well in the case of no sorting, significantly better in the case of sorting on common ability, and worse in the case of sorting on subject-specific ability. The latter is consistent with the identification assumption of the model.

I test the method using French administrative data for the universe of public middle schools in Metropolitan France. I focus on 9th graders and their Math and French teachers, exploiting the fact that these students need to hold standardised exams in the two subjects at the end of the 9th grade. I find that for a 1 SD increase in Math (French) teacher value added within school, student scores improve by 17.4 (16.3) percent of a SD. This implies that moving a student from a teacher at the 5th percentile of the value-added distribution of a school to one at the 95th percentile in the same school is associated with an average increase in student test scores by 58 (54) percent of a SD in Math (French).

I show that the variation of value added is higher in more disadvantaged regions of France and schools which have a higher share of disadvantaged students. Finally, I show that the teacher estimates are positively correlated with being female and with in-class pedagogical assessments, consistent with the literature. These results provide the first test of value added in French middle schools and indeed one of the first non-experimental evidence in Europe.

References

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

Bacher-Hicks, A., & Koedel, C. (2022). Chapter 2 - Estimation and interpretation of teacher value added in research applications. *Handbook of the Economics of Education*, *6*, 93–134.

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).

Buhl-Wiggers, J., Kerwin, J. T., Smith, J. A., & Thornton, R. (2017). The impact of teacher effectiveness on student learning in Africa .

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).

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, *40*, 76–92.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*, *10*(1), 117–156.

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

Jackson, C. K. (2014). Teacher Quality at the High School Level: The Importance of Accounting for Tracks. *Journal of Labor Economics*, *32*(4).

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, *26*(1).

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).

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill and Melinda Gates Foun- dation, Seattle, WA.*.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, *28*(6), 682–692.

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.

Lefgren, L., & Sims, D. (2012). Using Subject Test Scores Efficiently to Predict Teacher Value-Added. *Educational Evaluation and Policy Analysis*, *34*(1), 109–121.

Pommiers, E., & Lecherbonnier, S. (2022). Education : comment le système de mutation des enseignants s'est grippé. *Le Monde*. Retrieved from `https://www.lemonde.fr/societe/article/2022/09/05/mutation-des-enseignants-je-n-aurais-jamais-imagine-etre-encore-a-creteil-plus-de-dix-ans-apres_6140299_3224.html#:~:text=Les%20demandes%20de%20mutation%20entre,par%20d'importantes%20disparit%C3%A9s%20territoriales.`

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.

Appendix

A Deriving the identifying assumptions

A.1 Classroom size and multiple network observations

I turn to a case where we observe multiple classrooms within a school. Without loss of generality, let us assume there are four classrooms $c \in \{c_A, c_B, c_C, c_D\}$. For simplicity, I still focus on a case with a single direct network of Math teachers $j_M \in \{j_{M_{AB}}, j_{M_{CD}}\}$, linked through a single French teacher j_F who teaches in all four classrooms. For the sake of example, let $j_{M_{AB}}$ teach in classrooms c_A and c_B and $j_{M_{CD}}$ teach in classrooms c_C and c_D.

There are four possible ways to find the relative value added of teachers $j_{M_{AB}}$ and $j_{M_{CD}}$, using equation 4 and the decomposition of student ability $\gamma_{i,f}$:

$$\begin{aligned} &\Delta\widehat{A}_{M,c_A,c_C} - (\widehat{\overline{\mathbf{X}}}_{c_A} - \widehat{\overline{\mathbf{X}}}_{c_C}) \cdot (\beta_M - \beta_F) = \\ &(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + (\theta_{c_A,M} - \theta_{c_C,M}) - (\theta_{c_A,F} - \theta_{c_C,F}) \end{aligned} \tag{A.1}$$

$$\begin{aligned} &\Delta\widehat{A}_{M,c_A,c_D} - (\widehat{\overline{\mathbf{X}}}_{c_A} - \widehat{\overline{\mathbf{X}}}_{c_D}) \cdot (\beta_M - \beta_F) = \\ &(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + (\theta_{c_A,M} - \theta_{c_D,M}) - (\theta_{c_A,F} - \theta_{c_D,F}) \end{aligned} \tag{A.2}$$

$$\begin{aligned} &\Delta\widehat{A}_{M,c_B,c_C} - (\widehat{\overline{\mathbf{X}}}_{c_B} - \widehat{\overline{\mathbf{X}}}_{c_C}) \cdot (\beta_M - \beta_F) = \\ &(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + (\theta_{c_B,M} - \theta_{c_C,M}) - (\theta_{c_B,F} - \theta_{c_C,F}) \end{aligned} \tag{A.3}$$

$$\begin{aligned} &\Delta\widehat{A}_{M,c_B,c_D} - (\widehat{\overline{\mathbf{X}}}_{c_B} - \widehat{\overline{\mathbf{X}}}_{c_D}) \cdot (\beta_M - \beta_F) = \\ &(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + (\theta_{c_B,M} - \theta_{c_D,M}) - (\theta_{c_B,F} - \theta_{c_D,F}) \end{aligned} \tag{A.4}$$

I denote for simplicity

$$\Delta\widehat{A}_{M,c,c'} - (\widehat{\overline{\mathbf{X}}}_{c} - \widehat{\overline{\mathbf{X}}}_{c'}) \cdot (\beta_M - \beta_F) \equiv \Delta\widehat{A}^P_{M,c,c'}$$

and

$$[(\theta_{c,M} - \theta_{c',M}) - (\theta_{c,F} - \theta_{c',F})] \equiv \phi_{M,c,c'}$$

Equal classroom size Assuming that each classroom observation is equally informative of a teacher's value added due to equal classroom size, it follows for the average pairwise difference $(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}})$ that:

$$\begin{aligned} &\frac{1}{4}\sum_{c,c'} \widehat{A}^P_{M,c,c'} = \frac{1}{4}\sum_{c,c'}[(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \phi_{M,c,c'})] \implies \\ &\frac{1}{4}\sum_{c,c'} \widehat{A}^P_{M,c,c'} = (\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \frac{1}{4}\sum_{c,c'} \phi_{M,c,c'} \end{aligned} \tag{A.5}$$

for each $c \in \{c_A, c_B\}$ and $c' \in \{c_C, c_D\}$. The estimated difference in value added would then be equal to the true difference if:

$$\sum_{c,c'} \phi_{M,c,c'} = 0 \tag{A.6}$$

which is equivalent to:

$$\left[\sum_{c \in \{c_0 | j(c_0) = j_{M_{AB}}\}} (\theta_{c,M} - \theta_{c,F}) - \sum_{c' \in \{c'_0 | j(c'_0) = j_{M_{CD}}\}} (\theta_{c',M} - \theta_{c',F}) \right] = 0 \tag{A.7}$$

It follows that for the relative value added to be unbiased, the following condition needs to hold:

Condition.

$$\left[\sum_{c \in \{c_0 | j(c_0) = j_{M_{AB}}\}} (\theta_{c,M} - \theta_{c,F}) - \sum_{c' \in \{c'_0 | j(c'_0) = j_{M_{CD}}\}} (\theta_{c',M} - \theta_{c',F}) \right] = 0 \tag{A.8}$$

The Condition is satisfied if the average relative ability in Math compared to French for teacher $j_{M_{AB}}$ is the same as the average relative ability in Math compared to French for teacher $j_{M_{CD}}$. In other words, Math teachers should not be sorted on the relative ability in Math compared to French (and vice versa for French teachers). This leads us to the first assumption of the model.

Different classroom size So far I have assumed that each classroom observation is equally informative of a teacher's value added due to equal classroom size. Assume now that larger classrooms might are more informative due to a smaller noise in the estimated $\Delta \widehat{A}^P_{M,c,c'}$.

Classroom	**Classroom size**
c_A	30
c_B	10
c_C	20
c_D	20

Table A.1: Imaginary different classroom size

Consider that each of the four classrooms has a size as indicated in Table A.1. We can add different weights to each classroom accordingly, such that equation A.5 becomes:

$$\begin{aligned} &\frac{3}{8}\widehat{A}^P_{M,c_A,c_C} + \frac{1}{8}\widehat{A}^P_{M,c_A,c_D} + \frac{1}{4}\widehat{A}^P_{M,c_B,c_C} + \frac{1}{4}\widehat{A}^P_{M,c_B,c_D} = \\ &(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \frac{3}{8}\phi_{M,c_A,c_C} + \frac{1}{8}\phi_{M,c_A,c_D} + \frac{1}{4}\phi_{M,c_B,c_C} + \frac{1}{4}\phi_{M,c_B,c_D} \end{aligned} \tag{A.9}$$

It follows that for the estimated difference in value added to be equal to the true difference:

$$\frac{3}{8}\phi_{M,c_A,c_C} + \frac{1}{8}\phi_{M,c_A,c_D} + \frac{1}{4}\phi_{M,c_B,c_C} + \frac{1}{4}\phi_{M,c_B,c_D} = 0 \tag{A.10}$$

which can be expressed as:

$$\begin{aligned}&[\frac{1}{2}(\theta_{c_A,M} - \theta_{c_A,F}) + \frac{1}{2}(\theta_{c_B,M} - \theta_{c_B,F})] - \\ &[\frac{5}{8}(\theta_{c_C,M} - \theta_{c_C,F}) + \frac{3}{8}(\theta_{c_D,M} + \theta_{c_D,F})] = 0\end{aligned}$$

It follows that for the relative value added to be unbiased, the following condition needs to hold:

Condition.

$$\begin{aligned}&[w_{c_A}(\overline{\zeta}_{c_A,M} - \overline{\zeta}_{c_A,F}) + w_{c_B}(\overline{\zeta}_{c_B,M} - \overline{\zeta}_{c_B,F})] - \\ &[w_{c_C}(\overline{\zeta}_{c_C,M} - \overline{\zeta}_{c_C,F}) + w_{c_D}(\overline{\zeta}_{c_D,M} + \overline{\zeta}_{c_D,F})] = 0\end{aligned} \tag{A.11}$$

The Condition is satisfied if the weighted average relative ability in Math compared to French for teacher $j_{M_{AB}}$ is the same as the weighted average relative ability in Math compared to French for teacher $j_{M_{CD}}$.

This condition is very unlikely to hold in reality as it is highly specific. This indeed requires me to assume that each classroom is equally informative of a teacher's effectiveness, irrespective of differences in classroom size.

A.2 Drifts in value added

Case with only two classrooms Let us focus on a case where we observe only two classrooms of equal size within a school, but at different times $\tau \in \{t, t'\}$, such that $c \in \{c_{A,t}, c_{B,t'}\}$. Let us further assume that we observe two Math teachers, $j_{M,\tau} \in \{j_{M_A,t}, j_{M_B,t'}\}$, who teach respectively in classrooms $c_{A,t}$ and $c_{B,t'}$, and one French teacher j_F who teaches in both classrooms. This generalises the initially proposed model to:

$$\overline{A}_{c,f,\tau} = \overline{\mathbf{X}}_c\beta_f + \gamma_{c,f} + \theta_{j_f,\tau} + \theta_s + \varepsilon_{c,f,\tau}$$

where $\theta_{j_f,\tau} = \theta_{j_f} + \alpha_{j_f,\tau}$ and $\alpha_{j_f,\tau}$ is the time-specific part of a teacher's value added, which is additively separable from the intrinsic value added of a teacher. Then, following the same procedure as before, we can express for classroom A:

$$\begin{aligned}&\overline{A}_{c_{A,t},M,t} - \overline{A}_{c_{A,t},F,t} = \overline{\mathbf{X}}_{c_{A,t}} \cdot (\beta_M - \beta_F) + (\theta_{j_{M_A}} - \theta_{j_F}) + (\alpha_{j_{M_A},t} - \alpha_{j_F,t}) + \\ &(\theta_{c_{A,t},M} - \theta_{c_{A,t},F}) + (\varepsilon_{c_{A,t},M,t} - \varepsilon_{c_{A,t},F,t})\end{aligned}$$

and for classroom B:

$$\begin{aligned}&\overline{A}_{c_{B,t'},M,t'} - \overline{A}_{c_{B,t'},F,t'} = \overline{\mathbf{X}}_{c_{B,t'}} \cdot (\beta_M - \beta_F) + (\theta_{j_{M_B}} - \theta_{j_F}) + (\alpha_{j_{M_B},t'} - \alpha_{j_F,t'}) + \\ &(\theta_{c_{B,t'},M} - \theta_{c_{B,t'},F}) + (\varepsilon_{c_{B,t'},M,t'} - \varepsilon_{c_{B,t'},F,t'})\end{aligned}$$

Thus, one can express the relative value added of the two Math teachers as:

$$(\overline{A}_{c_{A,t},M,t} - \overline{A}_{c_{A,t},F,t}) - (\overline{A}_{c_{B,t'},M,t'} - \overline{A}_{c_{B,t'},F,t'}) = (\overline{\mathbf{X}}_{c_{A,t}} - \overline{\mathbf{X}}_{c_{B,t'}}) \cdot (\beta_M - \beta_F) +$$
$$(\theta_{j_{M_A}} - \theta_{j_{M_B}}) + [(\alpha_{j_{M_A},t} - \alpha_{j_{M_B},t'}) - (\alpha_{j_F,t} - \alpha_{j_F,t'})] +$$
$$(\theta_{c_{A,t},M} - \theta_{c_{A,t},F}) - (\theta_{c_{B,t'},M} - \theta_{c_{B,t'},F}) + (\varepsilon_{c_{A,t},M,t} - \varepsilon_{c_{A,t},F,t}) - (\varepsilon_{c_{B,t'},M,t'} - \varepsilon_{c_{B,t'},F,t'})$$

Assuming that Assumption 1 of the model holds and denoting for simplicity:

$$(\widehat{\overline{A}}_{c_{A,t},M,t} - \widehat{\overline{A}}_{c_{A,t},F,t}) - (\widehat{\overline{A}}_{c_{B,t'},M,t'} - \widehat{\overline{A}}_{c_{B,t'},F,t'}) - (\widehat{\overline{\mathbf{X}}}_{c_{A,t}} - \widehat{\overline{\mathbf{X}}}_{c_{B,t'}}) \cdot (\beta_M - \beta_F) \equiv \Delta \widehat{A}^{P'}_{M,c_{A,t},c_{B,t'}}$$

our estimated relative teacher value added can be expressed as:

$$\widehat{A}^{P'}_{M,c_{A,t},c_{B,t'}} = (\theta_{j_{M_A}} - \theta_{j_{M_B}}) + [(\alpha_{j_{M_A},t} - \alpha_{j_{M_B},t'}) - (\alpha_{j_F,t} - \alpha_{j_F,t'})]$$

The predicted relative value added of teachers j_{M_A} and j_{M_B} from this alternative model would be equal to that of my proposed model given the following condition:

Condition

$$\mathbb{E}[(\alpha_{j_{M_A},t} - \alpha_{j_{M_B},t'}) - (\alpha_{j_F,t} - \alpha_{j_F,t'})] = 0 \tag{A.12}$$

Note that this condition would hold under the stricter assumption $\mathbb{E}[\alpha_{j_f,\tau}] = 0$. However, I can instead assume that $\alpha_{j_f,\tau}$ is some functional form of years of experience, such that for Math, $\alpha_{j_M,\tau} = f(exp_{j_M,\tau})$ and $\alpha_{j_F,\tau} = g(exp_{j_F,\tau})$. Since a teacher's years of experience are observable, one can simply control for $f(exp_{j_M,\tau})$ and $g(exp_{j_F,\tau})$ in $\overline{\mathbf{X}}_{c_\tau}$.

Case with multiple classrooms I turn to a case where we observe multiple classrooms within a school. Let us assume there are four classrooms $c \in \{c_{A,t_1}, c_{B,t_2}, c_{C,t_3}, c_{D,t_4}\}$ observed in 4 different periods, $\tau \in [t_1, t_4]$. For simplicity, I still focus on a case with a single direct network of Math teachers $j_M \in \{j_{M_{AB}}, j_{M_{CD}}\}$, linked through a single French teacher j_F who teaches in all four classrooms. For the sake of example, let $j_{M_{AB}}$ teach in classrooms c_{A,t_1} and c_{B,t_2} and $j_{M_{CD}}$ teach in classrooms c_{C,t_3} and c_{D,t_4}.

As before, there are four possible ways to find the relative value added of teachers $j_{M_{AB}}$ and $j_{M_{CD}}$. Denoting for simplicity

$$\Delta \widehat{A}_{M,c,c'} - (\widehat{\overline{\mathbf{X}}}_c - \widehat{\overline{\mathbf{X}}}_{c'}) \cdot (\beta_M - \beta_F) \equiv \Delta \widehat{A}^{P'}_{M,c,c'}$$

for each $c \in \{c_{A,t_1}, c_{B,t_2}\}$ and $c' \in \{c_{C,t_3}, c_{D,t_4}\}$, and

$$(\alpha_{j_M,t} - \alpha_{j_{M'},t'}) - (\alpha_{j_F,t} - \alpha_{j_F,t'}) \equiv \Delta\alpha_{j_M,j_{M'},t,t'}$$

for each $t \in \{t_1, t_2\}$ and $t' \in \{t_3, t_4\}$, and furthermore assuming Assumptions 1 and 2 of the model hold, I can express these as:

$$\Delta \widehat{A}^{P'}_{M,c_{A,t_1},c_{C,t_3}} = (\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \Delta\alpha_{j_{M_{AB}},j_{M_{CD}},t_1,t_3} \tag{A.13}$$

$$\Delta\widehat{A}^{P'}_{M,c_{A,t_1},c_{D,t_4}} = (\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \Delta\alpha_{j_{M_{AB}},j_{M_{CD}},t_1,t_4} \tag{A.14}$$

$$\Delta\widehat{A}^{P'}_{M,c_{B,t_2},c_{C,t_3}} = (\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \Delta\alpha_{j_{M_{AB}},j_{M_{CD}},t_2,t_3} \tag{A.15}$$

$$\Delta\widehat{A}^{P'}_{M,c_{B,t_2},c_{D,t_4}} = (\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \Delta\alpha_{j_{M_{AB}},j_{M_{CD}},t_2,t_4} \tag{A.16}$$

Assuming that Assumption 3 of the model holds, it follows that for the average pairwise difference $(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}})$:

$$\begin{aligned}
&\frac{1}{4}\sum_{c,c'}\widehat{A}^{P'}_{M,c,c'} = \frac{1}{4}\sum_{c,c'}[(\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \Delta\alpha_{j_M,j_{M'},t,t'}] \implies \\
&\frac{1}{4}\sum_{c,c'}\widehat{A}^{P}_{M,c,c'} = (\theta_{j_{M_{AB}}} - \theta_{j_{M_{CD}}}) + \frac{1}{4}\sum_{c,c'}\Delta\alpha_{j_M,j_{M'},t,t'}
\end{aligned} \tag{A.17}$$

The estimated difference in value added would then be equal to the true difference if:

$$\mathbb{E}[\sum_{c,c'}\Delta\alpha_{j_M,j_{M'},t,t'}] = 0 \tag{A.18}$$

which can be rewritten as:

$$\mathbb{E}\left[\sum_{t,t'}(\alpha_{j_M,t} - \alpha_{j_{M'},t'}) - \sum_{t,t'}(\alpha_{j_F,t} - \alpha_{j_F,t'})\right] = 0 \tag{A.19}$$

Note that this holds as Condition A.12, given the functional forms of $\alpha_{j_f,\tau}$ for each subject f. Therefore, this leads me to the third assumption of the model.

B Tests of within-school teacher-student sorting

Student-teacher sorting may occur under two conditions. Firstly, if the school groups students into classrooms by ability (tracking). And secondly, if some teachers specialise into teaching either low or high tracks.

While one cannot test for tracking and specialisation based on unobservable student characteristics by default, we could test for sorting with the use of observable characteristics. In particular, if there is tracking and specialisation, then this should be detectable from the observable characteristics of students grouped together in a classroom and assigned to a teacher. One natural candidate for such characteristics is past achievement. However, as we do not observe any student scores prior to the *DNB* grades, a second-best alternative is to rely on other student observables.

I predict student grades as shown in Table B.2 for Math, French, and the difference between a student's grades in Math and French which represents the student's *relative Math grade*. I denote these predicted scores in Math and French for student i respectively $A^P_{i,M}$ and $A^P_{i,F}$, such that $A^P_{i,M-F}$ denotes the predicted relative Math grade of student i. The used observable characteristics are able to predict 30.4% of the variation in Math, 28.7% in French, and 11.4% in relative Math to French grade, according to the adjusted R^2.

Table B.2: Predicted grades

	Math grade	French grade	Math-French grade
Male	0.030***	-0.379***	0.409***
	(0.001)	(0.001)	(0.001)
Age	-0.364***	-0.333***	-0.032***
	(0.001)	(0.001)	(0.001)
Scholarship student	-0.141***	-0.114***	-0.027***
	(0.001)	(0.001)	(0.001)
High SES	-0.267***	-0.224***	-0.043***
	(0.002)	(0.002)	(0.001)
Medium SES	-0.404***	-0.351***	-0.053***
	(0.001)	(0.001)	(0.001)
Low SES	-0.618***	-0.582***	-0.036***
	(0.002)	(0.002)	(0.002)
Moved department	0.018***	0.037***	-0.019***
	(0.001)	(0.001)	(0.001)
Taking Ancient Greek	0.411***	0.399***	0.012***
	(0.003)	(0.003)	(0.003)
Taking Latin	0.540***	0.552***	-0.013***
	(0.001)	(0.001)	(0.001)
School FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Nationality FE	Yes	Yes	Yes
Main guardian type FE	Yes	Yes	Yes
Observations	2,807,201	2,807,201	2,807,201
Adjusted R^2	0.304	0.287	0.114

Note: The table presents the OLS regression results for the association of observable student characteristics with student grades in Math, French and the student-level difference between Math and French grades. Student grades are standardised by subject and year. The base SES category is that of "Very high SES". The dummy "Moved department" is equal to 1 if a student's birth department is different from his current department of living.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B.1 Tracking

I test for tracking of students into classrooms by regressing the predicted grades $A_{i,f} \in \{A^P_{i,M}, A^P_{i,F}, A^P_{i,M-F}\}$ on classroom fixed effects for each school s:

$$A^P_{i,f} = \theta_{c_n} + \varepsilon_{i,f}$$

As each classroom dummy captures the specific constant for each classroom, R^2 of the regression could be interpreted as a measure of between-classroom segregation in student characteristics in the specific school s. Repeating this regression for all schools in the system of schools with a complete school network (i.e. all the schools included in the analysis) and taking the average R^2 therefore provides us with an average measure of such between-classroom segregation. Let us denote this average percentage as the *True* R^2.

We can then reassign students to a random classroom numerous times and compute the R^2 from these simulations, which would give us a placebo measure of between-classroom segregation given random allocation of students to classrooms. In particular, I do this reassignment procedure 100 times, reassigning students only to a classroom from the universe of classrooms in their school and year of study, and making sure that the classroom size of each classroom c_n is kept fixed to its original value. Following the same procedure as above, we can take the average R^2 as the average measure of between-classroom segregation under random reassignment. Let us denote this average percentage as the *Placebo* R^2.

Using the two measures, we can then compute at which percentile p of the bootstrap distribution of the *Placebo* R^2 is the value of the *True* R^2. In this setting, $1-p$ can be interpreted as the *empirical p-value*: the probability that the observed level of segregation could have occurred by chance.

The first three sets of bars in Figure B.1 depict the *True* R^2, *Placebo* R^2 and *empirical p-value* for the predicted Math grade, the predicted French grade and the predicted relative Math grade for the tests of student tracking.

What we can see is that, while there is a small baseline segregation in student characteristics in Math, there is a very small probability (6%) that this segregation has occurred by chance. The same conclusions follow for French (7% probability). However, once we turn to relative Math, we see that the level of the baseline segregation in student characteristics falls significantly and furthermore that the probability that this segregation has occurred by chance jumps to 73%.

B.2 Teacher specialisation

I perform similar tests to determine the presence of teacher sorting to classrooms within a school, particularly by regressing the students' predicted grades in a subject on teacher fixed effects:

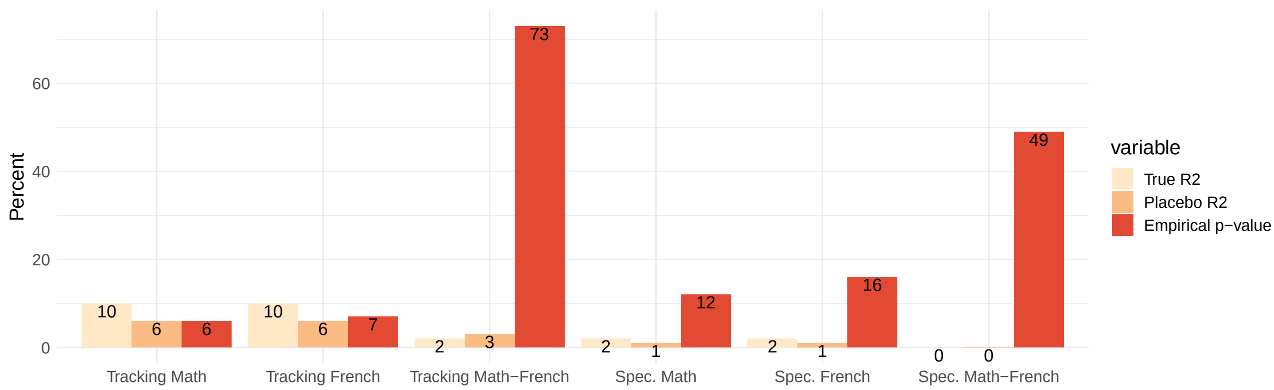
$$A^P_{i,f} = \theta_j + \varepsilon_{i,f}$$

In this setting, the *True* R^2 now represents the average measure of between-teacher segregation in student characteristics, which would indicate that different teachers specialise in teaching different types of students. The *Placebo* R^2 is the corresponding average measure of between-teacher segregation under random reassignment of teachers to students. Similarly, the *empirical p-value* shows the probability that the observed level of segregation could have occurred by chance.

The last three sets of bars in Figure B.1 depict the *True* R^2, *Placebo* R^2 and *empirical p-value* for the predicted Math grade, the predicted French grade and the predicted relative Math grade for the tests of teacher specialisation.

Figure B.1: Tests of within-school teacher-student non-random sorting

Note: The figure shows the results from the tests on student tracking (the 3 sets of columns on the left) and specialisation (the 3 columns on the right). For both tests, I predict student grades in Math, French and the difference between Math and French grade (relative Math grade) on observable student characteristics, as shown in Table B.2. Using the predicted grades, I first test for tracking (specialisation) by regressing each of the predicted grades on classroom (teacher) fixed effects. I take the R^2 of this regression as the measure of between-classroom (teacher) segregation in school s. The average R^2 across all schools is denoted as the *True* R^2 in this figure. I then reassign students to random classrooms (teachers) 100 times, keeping classroom size fixed to its original value, and compute the R^2 from these simulations. I denote this the *Placebo* R^2, and it is interpreted as a measure of between-classroom (teacher) segregation under random allocation. I compute at which percentile p of the bootstrap distribution of the *Placebo* R^2 is the value of the *True* R^2. The *Empirical p-value* is $1 - p$, the probability that the observed level of segregation could have occurred by chance.



The conclusions from the previous section follow similarly here. In fact, segregation of student characteristics in Math and in French are even smaller for segregation across teachers than it is for segregation across classrooms. Furthermore, once we turn to relative Math, the baseline level of between-teacher segregation is basically zero, with a probability that this segregation has occurred by chance of 49%.

Therefore, taking into account both the results for tracking and specialisation, and assuming that similar conclusions would follow for unobservable student characteristics, it is not difficult to argue in favour of the proposed methodology, as the method requires solely that student-teacher sorting does not occur based on the *relative Math ability* of students (or other such unobservable factors).

C Tests of differential classroom informativeness of teacher effects

In this section, I focus on the handling of multiple observations of a network of two teachers. In particular, one might be worried that larger classrooms are more informative of a teacher's value added. To test for this, I consider a network of two teachers which is observed twice - through different link teachers, as shown in Table C.3.

Math teacher	**French teacher**	**Classroom**	**Classroom size**
$j_{M_{AB}}$	$j_{F_{AC}}$	c_A	30
$j_{M_{AB}}$	$j_{F_{BD}}$	c_B	10
$j_{M_{CD}}$	$j_{F_{AC}}$	c_C	20
$j_{M_{CD}}$	$j_{F_{BD}}$	c_D	20

Table C.3: Imaginary network of teachers with different classroom size

As one can see, through their connection to the link teacher $j_{F_{AC}}$, teachers $j_{M_{AB}}$ and $j_{M_{CD}}$ share a total of 50 students, whereas through their connection to the link teacher $j_{F_{BD}}$ they share only 30 students.[29,30] I can reflect the fact that for the former, my estimation strategy would lead to more precise estimates than for the latter by adding weights. In particular, following the empirical strategy detailed in Section 4, the second-stage equations for the network of Math teachers would be:

$$(\widehat{\varepsilon}_{c_A,j_{M_{AB}}} - \widehat{\varepsilon}_{c_A,j_{F_{AC}}}) - (\widehat{\varepsilon}_{c_C,j_{M_{CD}}} - \widehat{\varepsilon}_{c_C,j_{F_{AC}}}) = \Delta_1 \tag{C.20}$$

and

$$(\widehat{\varepsilon}_{c_B,j_{M_{AB}}} - \widehat{\varepsilon}_{c_B,j_{F_{BD}}}) - (\widehat{\varepsilon}_{c_D,j_{M_{CD}}} - \widehat{\varepsilon}_{c_D,j_{F_{BD}}}) = \Delta_2 \tag{C.21}$$

[29] Note that in reality, the differences in classroom size are much smaller, and extremely small classrooms (below 15 students) are removed from the analysis due to likely being reporting errors, so as to minimise the noise in estimation even further.

[30] Conclusions hold in the same way given different observations of a network with the same link teacher.

I can compute $\widehat{\theta}_{j_{M_{AB}}} - \widehat{\theta}_{j_{M_{CD}}}$ as:

$$\widehat{\theta}_{j_{M_{AB}}} - \widehat{\theta}_{j_{M_{CD}}} = \frac{5}{8}\Delta_1 + \frac{3}{8}\Delta_2 \tag{C.22}$$

Following this logic, I use the estimated residuals to compute for each observation of a network of two Math teachers:

$$(\widehat{\varepsilon}_{c_A,j_{M_A}} - \widehat{\varepsilon}_{c_A,j_F}) - (\widehat{\varepsilon}_{c_B,j_{M_B}} - \widehat{\varepsilon}_{c_B,j_F}) = \Delta^n_{j_{M_A},j_{M_B}}$$

and similarly for a network of two French teachers:

$$(\widehat{\varepsilon}_{c_C,j_{F_C}} - \widehat{\varepsilon}_{c_C,j_M}) - (\widehat{\varepsilon}_{c_D,j_{F_D}} - \widehat{\varepsilon}_{c_D,j_M}) = \Delta^n_{j_{F_C},j_{F_D}}$$

I then average across all observations of a single network of two teachers in a subject, weighting by the number of students within each observation, such that the relative value-added of the Math teachers j_{M_A} and j_{M_B} is computed as:

$$\widehat{\theta}_{j_{M_A}} - \widehat{\theta}_{j_{M_B}} = \sum_1^N w_n \cdot \Delta^n_{j_{M_A},j_{M_B}}$$

and similarly that of the French teachers j_{F_C} and j_{F_D} is:

$$\widehat{\theta}_{j_{F_C}} - \widehat{\theta}_{j_{F_D}} = \sum_1^N w_n \cdot \Delta^n_{j_{F_C},j_{F_D}}$$

where w is the weight of observation n.

I apply these methods within school and calculate the standard deviation of value-added measured within each school. I then proceed by calculating the average of these within-school standard deviations, in order to provide an estimate of the average within-school variability in teacher value-added.

Comparing the results of this estimation to the results of the baseline estimation, I note two facts. Firstly, the Spearman correlation between individual value-added estimates is 0.99 in for both Math and French within-school value-added distributions. Secondly, the estimated average within-school standard deviation in value added is almost identical: for Math, the new s.d. is 0.1743 compared to the baseline value of 0.1744; similarly, for French, the new s.d. is 0.1632 compared to 0.1633. Therefore, I conclude that adding weights based on classroom size does not add much value to the analysis.

D Simulation exercise

D.1 Model comparison

Network estimates In line with the theoretical model described in my paper, as $A^s_{i,f,j_f,t}$ represent the student grades in each subject f, net of student characteristics, it follows that:

$$(A^s_{c_A,M,j_{M_A}} - A^s_{c_A,M,j_F}) - (A^s_{c_B,M,j_{M_B}} - A^s_{c_B,M,j_F}) \equiv A^{s,n}_{j_{M_A},j_{M_B}} \tag{D.23}$$

and therefore, using all n observations of the network of teachers j_{M_A} and j_{M_B}, I find their relative value added as:

$$\widehat{\theta}_{j_{M_A}} - \widehat{\theta}_{j_{M_B}} = \frac{1}{N}\sum_{1}^{N} A^{s,n}_{j_{M_A},j_{M_B}} \tag{D.24}$$

I use the method described in Section 4 to move from pairwise comparisons to a distribution of value added.

Note that according to my data, the correlation between the residualised Math and French grades of students is 0.58. I take this value as a reference point but also vary the level of correlation between the two grades (through an increase in the standard deviation of the French grade) in order to see how the MSE and rank predictability react.

In the case of no sorting between students and teachers, both θ_i and $\theta_{i,M}$ are left to vary independently of θ_{j_M}.

Once sorting on common ability is introduced, I use an extreme example such that θ_i is 1-to-1 correlated with teacher value added θ_{j_M}, in the form of $\theta_i = a \times \theta_{j_M} + b$, $a > 0$. This would mean that better Math teachers are sorted into classrooms with better students., It is clear that, using equation 7 to substitute into equation D.23, this factor would cancel out in the difference.

Finally, in the case of sorting on Math-specific ability, to use an extreme example, let $\theta_{i,M}$ now be 1-to-1 correlated with teacher value added θ_{j_M}, such that $\theta_{i,M} = a \times \theta_{j_{j_M}} + b$, $a > 0$. Using again equation 7 averaged on a classroom level to substitute into equation D.23, and assuming there is no sorting in French such that $\theta_{c_A,F} = \theta_{c_B,F}$ on average, then:

$$A^{s,n}_{j_{M_A},j_{M_B}} = (1+a)[\theta_{j_{M_A}} - \theta_{j_{M_B}}] \neq \theta_{j_{M_A}} - \theta_{j_{M_B}} \tag{D.25}$$

In both cases, to make sure that the transformation does not change the relative part of the variance of the simulated grade, I then restandardise respectively the θ_i in the first case and $\theta_{i,M}$ in the second, to have the same standard deviation as in the no sorting case.

Baseline estimates I reproduce the BE by running a regression of the sort:

$$A^{s}_{i,M,j_M} = \alpha + \beta A_{i,M,j_{M'},t-1} + \theta_{j_M} + \epsilon_{i,M,j_M} \tag{D.26}$$

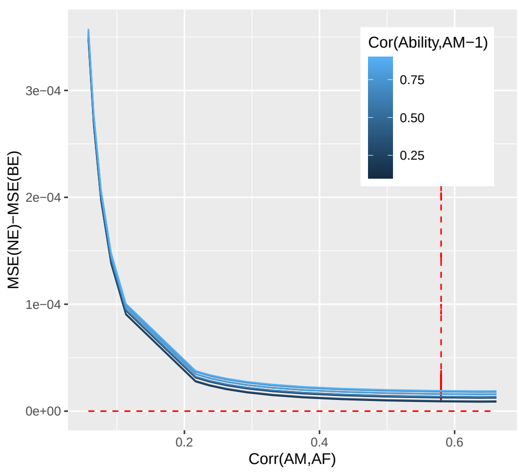
and extract the estimated teacher fixed effects θ_{j_M}. Rothstein (2009) shows that the correlation for actual reading grades between classes is between 0.7 and 0.8. I take 0.8 as a reference point for the correlation between the ability of a student $(\theta_i + \theta_{i,M})$ and his past Math grade, ρ, but I also repeat the exercise for different values of ρ in order to see how the MSE and rank predictability react. Note that we do not have an idea of ρ from the literature, as it represents the correlation between *residualised* grades, rather than actual grades, but it is logical that such correlation would be lower, as a lot of the same observable student-level factors affect the realisation of both grades.

In this case, for the case of any of the two types of sorting, as the correlation between $A_{i,M,j_{M'},t-1}$ and $(\theta_i + \theta_{i,M})$ is lower than 1, the use of the past grade does not fully clear up the bias of sorting.

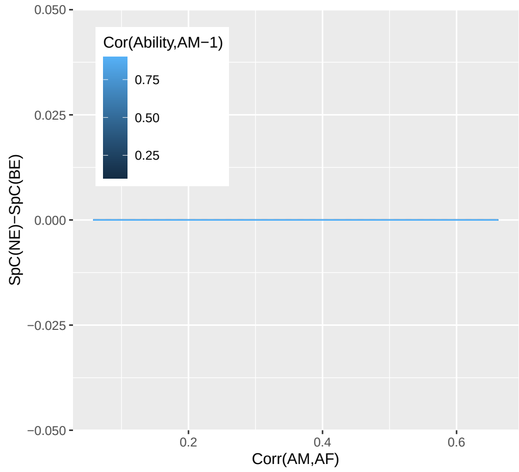
Figure D.2: Results for a case of random teacher-student sorting

The four graphs represent the results for the BE and NE models in the case of no sorting. Panels (a) and (b) depict respectively the difference in the mean squared error (MSE) of the NE and BE models, and the difference between the Spearman correlation of estimates of the NE and BE models, when comparing the value added estimates of each to the true value added parameters. In both graphs, the x-axis represents variation in the correlation between $A^s_{i,M}$ and $A^s_{i,F}$, such that the vertical red line represents the empirically found correlation of 0.58. Each of the lines represents a different level of correlation between student ability $\theta_i + \theta_{i,M}$ and past grade $A_{i,M,t-1}$. The crossing of the horizontal red line is associated with the NE method outperforming the BE method in terms of precision of estimates. Panels (c) and (d) depict respectively the estimates of value added for the 5 Math teachers from both the NE and BE model, and how they compare to the true value added parameters (the 45-degree line). The estimates of the NE model are shown with circles and the estimates of the BE model are shown with triangles. Panel (c) keeps constant the corr($A_{M,t}$,$A_{M,t-1}$), at 0.8, and varies the corr(A_M,A_F), therefore showing a set of NE estimates for each value of corr(A_M,A_F). Panel (d) keeps constant the corr(A_M,A_F), at 0.58, and varies the corr($A_{M,t}$,$A_{M,t-1}$), therefore showing a set of BE estimates for each value of corr($A_{M,t}$,$A_{M,t-1}$).

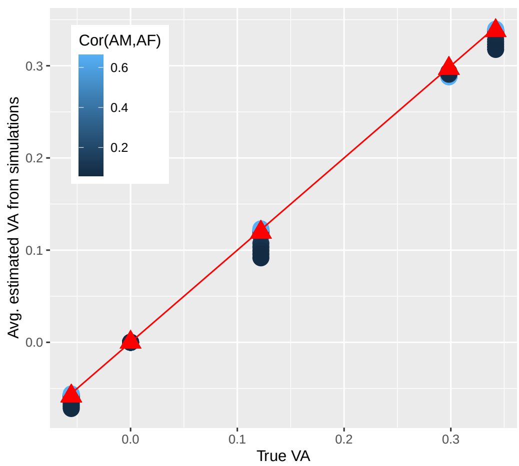
(a) Δ MSE of NE and MSE of BE



(b) Δ Spearman corr. of NE and of BE



(c) Distribution of NE VA for different level of corr(A_M,A_F) and BE VA for corr($A_{M,t}$,$A_{M,t-1}$)=0.8



(d) Distribution of BE VA for different level of corr($A_{M,t}$,$A_{M,t-1}$) and NE VA for corr(A_M,A_F)=0.58

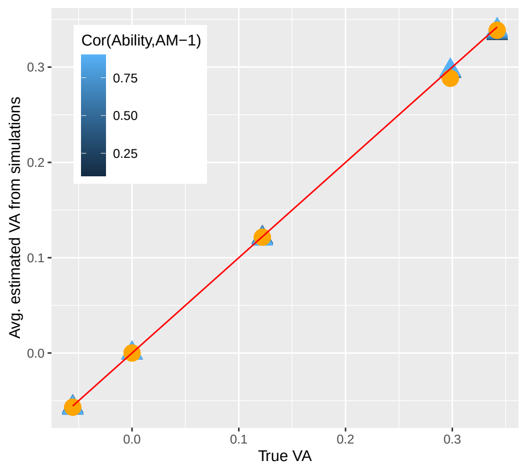
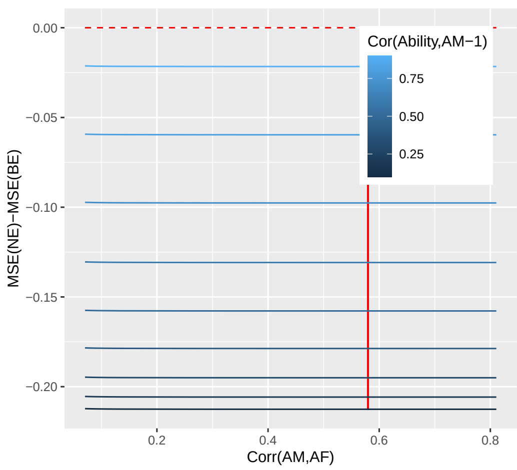


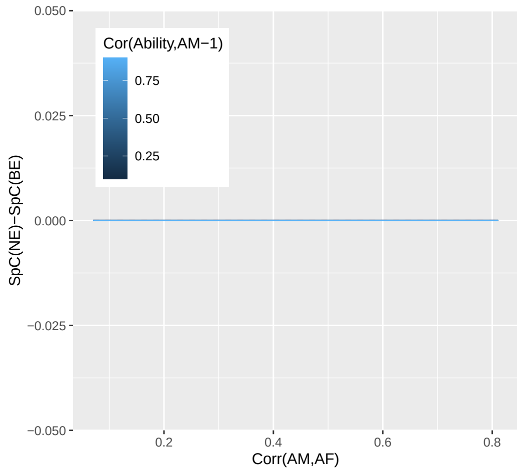
Figure D.3: Results for a case of teacher-student sorting based on common ability

The four graphs represent the results for the BE and NE models in the case of no sorting. Panels (a) and (b) depict respectively the difference in the mean squared error (MSE) of the NE and BE models, and the difference between the Spearman correlation of estimates of the NE and BE models, when comparing the value added estimates of each to the true value added parameters. In both graphs, the x-axis represents variation in the correlation between $A^s_{i,M}$ and $A^s_{i,F}$, such that the vertical red line represents the empirically found correlation of 0.58. Each of the lines represents a different level of correlation between student ability $\theta_i + \theta_{i,M}$ and past grade $A_{i,M,t-1}$. The crossing of the horizontal red line is associated with the NE method outperforming the BE method in terms of precision of estimates. Panels (c) and (d) depict respectively the estimates of value added for the 5 Math teachers from both the NE and BE model, and how they compare to the true value added parameters (the 45-degree line). The estimates of the NE model are shown with circles and the estimates of the BE model are shown with triangles. Panel (c) keeps constant the corr($A_{M,t}$,$A_{M,t-1}$), at 0.8, and varies the corr(A_M,A_F), therefore showing a set of NE estimates for each value of corr(A_M,A_F). Panel (d) keeps constant the corr(A_M,A_F), at 0.58, and varies the corr($A_{M,t}$,$A_{M,t-1}$), therefore showing a set of BE estimates for each value of corr($A_{M,t}$,$A_{M,t-1}$).

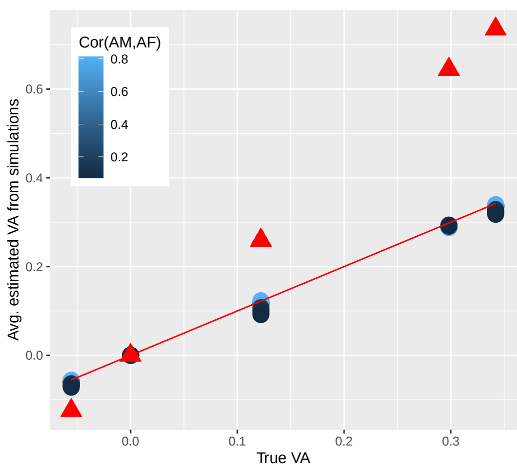
(a) Δ MSE of NE and MSE of BE



(b) Δ Spearman corr. of NE and of BE



(c) Distribution of NE VA for different level of corr(A_M,A_F) and BE VA for corr($A_{M,t}$,$A_{M,t-1}$)=0.8



(d) Distribution of BE VA for different level of corr($A_{M,t}$,$A_{M,t-1}$) and NE VA for corr(A_M,A_F)=0.58

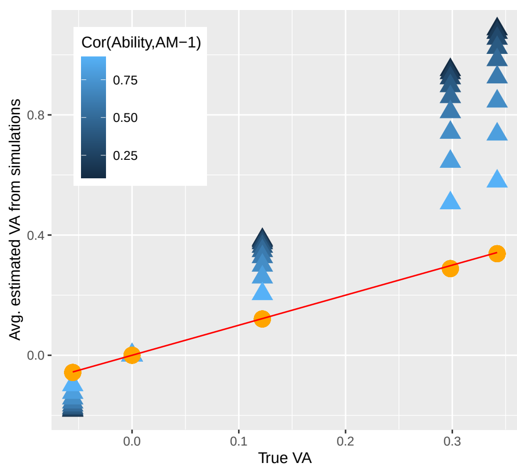
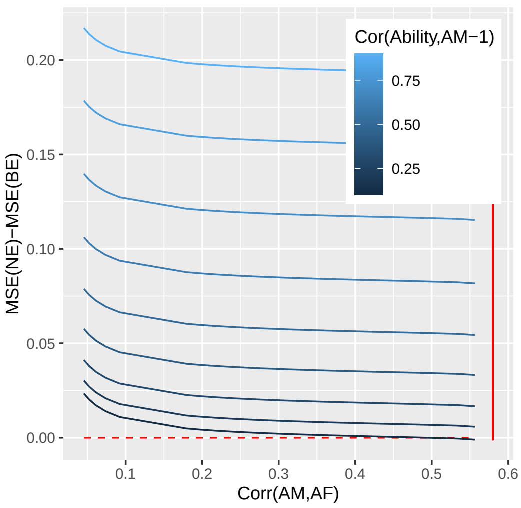


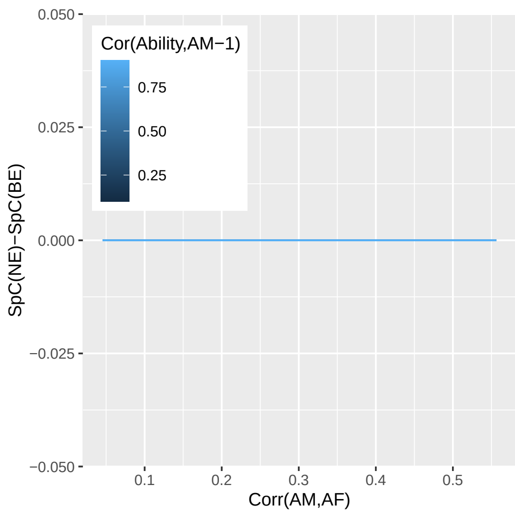
Figure D.4: Results for a case of teacher-student sorting based on subject-specific ability

The four graphs represent the results for the BE and NE models in the case of no sorting. Panels (a) and (b) depict respectively the difference in the mean squared error (MSE) of the NE and BE models, and the difference between the Spearman correlation of estimates of the NE and BE models, when comparing the value added estimates of each to the true value added parameters. In both graphs, the x-axis represents variation in the correlation between $A^s_{i,M}$ and $A^s_{i,F}$, such that the vertical red line represents the empirically found correlation of 0.58. Each of the lines represents a different level of correlation between student ability $\theta_i + \theta_{i,M}$ and past grade $A_{i,M,t-1}$. The crossing of the horizontal red line is associated with the NE method outperforming the BE method in terms of precision of estimates. Panels (c) and (d) depict respectively the estimates of value added for the 5 Math teachers from both the NE and BE model, and how they compare to the true value added parameters (the 45-degree line). The estimates of the NE model are shown with circles and the estimates of the BE model are shown with triangles. Panel (c) keeps constant the corr($A_{M,t}$,$A_{M,t-1}$), at 0.8, and varies the corr(A_M,A_F), therefore showing a set of NE estimates for each value of corr(A_M,A_F). Panel (d) keeps constant the corr(A_M,A_F), at 0.58, and varies the corr($A_{M,t}$,$A_{M,t-1}$), therefore showing a set of BE estimates for each value of corr($A_{M,t}$,$A_{M,t-1}$).

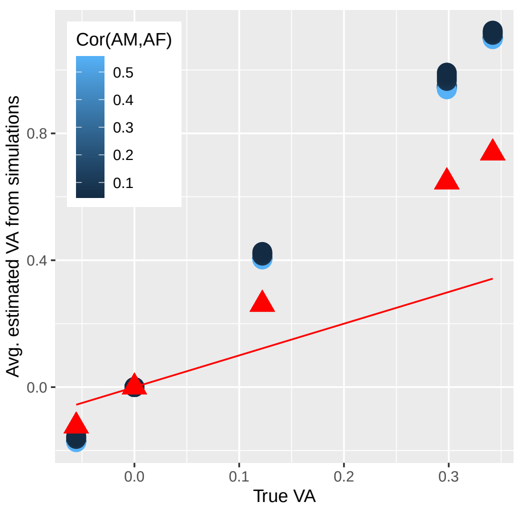
(a) Δ MSE of NE and MSE of BE



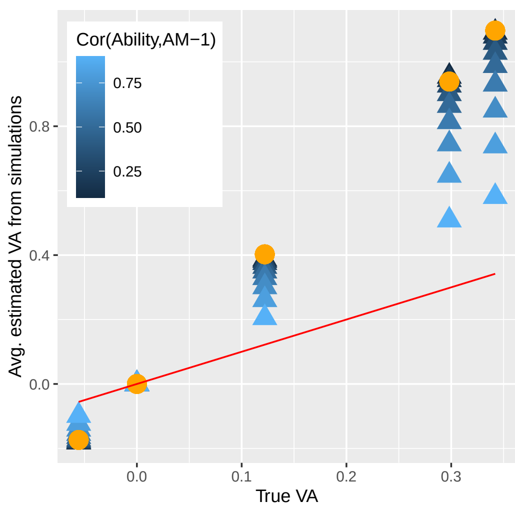
(b) Δ Spearman corr. of NE and of BE



(c) Distribution of NE VA for different level of corr(A_M,A_F) and BE VA for corr($A_{M,t}$,$A_{M,t-1}$)=0.8



(d) Distribution of BE VA for different level of corr($A_{M,t}$,$A_{M,t-1}$) and NE VA for corr(A_M,A_F)=0.58

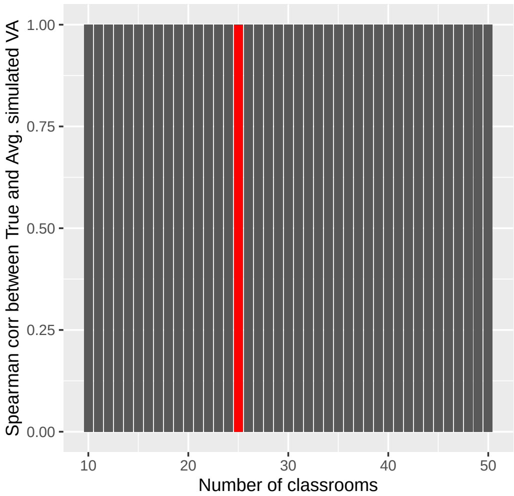


D.2 Robustness

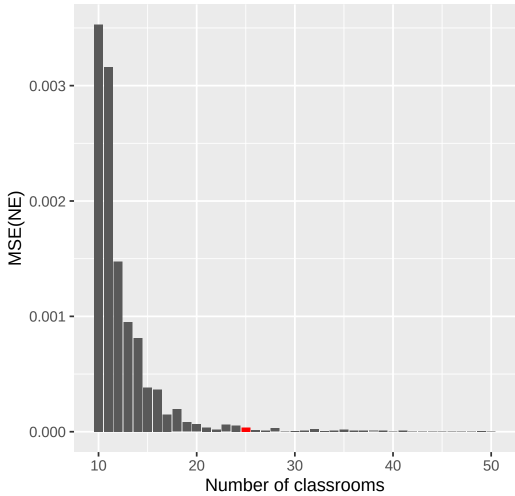
To show that the results outlined above are robust to changes in the characteristics of the school, I vary sequentially the number of classrooms, the average classroom size, and the number of Math and French teachers and plot the respective Spearman correlation and MSE for the NE model, assuming no sorting. Every time a characteristic is left to vary, I keep all other characteristics constant for tractability.

Figure D.5: Spearman correlation (LHS) and MSE (RHS) of the NE model for different levels of the school characteristics (pt. 1)

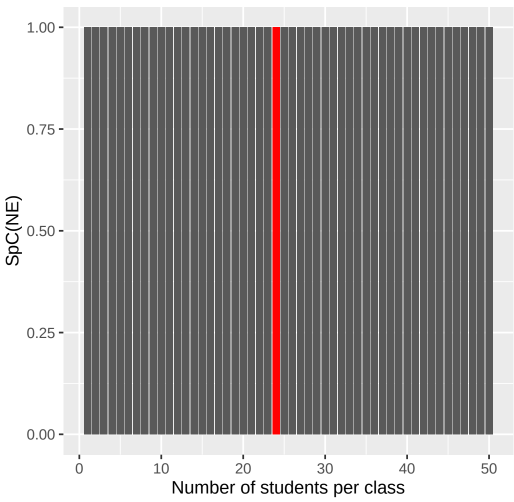
(a) Number of classrooms



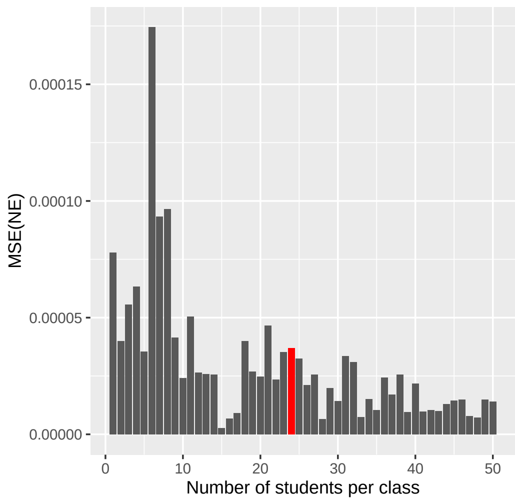
(b) Number of classrooms



(c) Classroom size



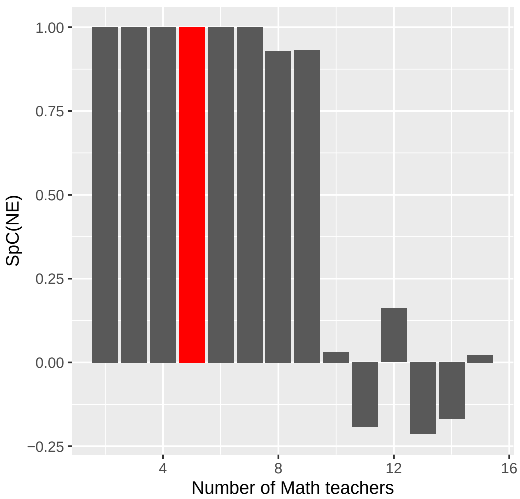
(d) Classroom size



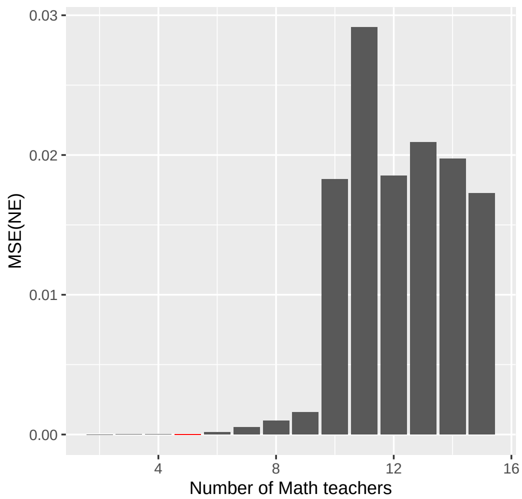
Overall, changes in the number of classrooms, the classroom size and the number of French (link) teachers does not affect the rank prediction of the model, but slightly affects the MSE, such that fewer classrooms lead to a larger MSE. The number of Math teachers (i.e. those for whom I estimate value added) naturally plays a larger part, such that too many teachers lead to too few observations per teacher in a classroom, and therefore leads to a low Spearman correlation and high MSE.

Figure D.6: Spearman correlation (LHS) and MSE (RHS) of the NE model for different levels of the school characteristics (pt. 2)

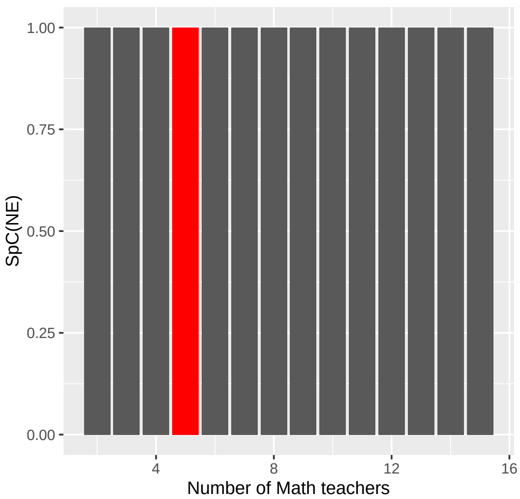
(a) Number of Math teachers



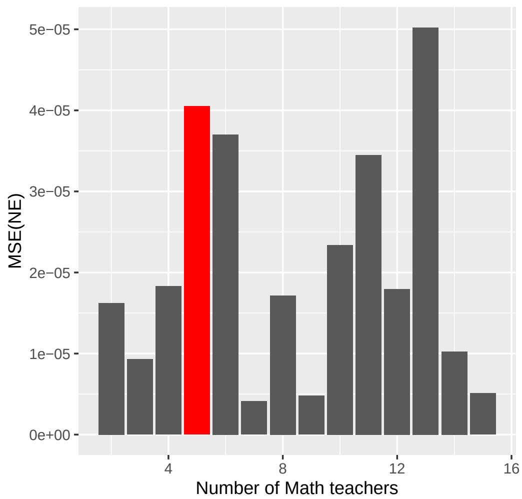
(b) Number of Math teachers



(c) Number of French (link) teachers



(d) Number of French (link) teachers



I further vary the respective shares in the total variance of overall ability, Math-specific ability and teacher value added.

	NE	BE	Δ (NE-BE)
No sorting			
Shares ($\theta_i = 0.23, \theta_{i,M} = 0.69, \theta_{j_M} = 0.07$)			
Sqr. Root of MSE	0.005	0.004	0.001
Spearman corr.	1.00	1.00	0.00
Shares ($\theta_i = 0.69, \theta_{i,M} = 0.23, \theta_{j_M} = 0.07$)			
Sqr. Root of MSE	0.004	0.003	0.001
Spearman corr.	1.00	1.00	0.00
Shares ($\theta_i = 0.395, \theta_{i,M} = 0.395, \theta_{j_M} = 0.21$)			
Sqr. Root of MSE	0.006	0.004	0.002
Spearman corr.	1.00	1.00	0.00
Sorting on common ability			
Shares ($\theta_i = 0.23, \theta_{i,M} = 0.69, \theta_{j_M} = 0.07$)			
Sqr. Root of MSE	0.005	0.14	-0.135
Spearman corr.	1.00	1.00	0.00
Shares ($\theta_i = 0.69, \theta_{i,M} = 0.23, \theta_{j_M} = 0.07$)			
Sqr. Root of MSE	0.004	0.393	-0.389
Spearman corr.	1.00	1.00	0.00
Shares ($\theta_i = 0.395, \theta_{i,M} = 0.395, \theta_{j_M} = 0.21$)			
Sqr. Root of MSE	0.006	0.225	-0.219
Spearman corr.	1.00	1.00	0.00
Sorting on subject-specific ability			
Shares ($\theta_i = 0.23, \theta_{i,M} = 0.69, \theta_{j_M} = 0.07$)			
Sqr. Root of MSE	0.567	0.394	0.173
Spearman corr.	1.00	1.00	0.00
Shares ($\theta_i = 0.69, \theta_{i,M} = 0.23, \theta_{j_M} = 0.07$)			
Sqr. Root of MSE	0.329	0.142	0.187
Spearman corr.	1.00	1.00	0.00
Shares ($\theta_i = 0.395, \theta_{i,M} = 0.395, \theta_{j_M} = 0.21$)			
Sqr. Root of MSE	0.427	0.227	0.2
Spearman corr.	1.00	1.00	0.00

Table D.4: Model performance comparison

Note: This table depicts the square root of the mean squared error (MSE) of the Traditional model (BE) compared to the true value added parameters and of the Network model (NE) and the true value added parameters. It also shows the Spearman correlations between the BE model and the true value added parameters, and the NE model and the true value added parameters. The referenced BE model is estimated for a correlation $\rho = 0.8$ between past student grade and student ability ($\theta_i + \theta_{i,f}$). The referenced NE model is estimated for a correlation 0.58 between A^s_{i,M,j_M} and A^s_{i,F,j_F}. The case of *No sorting* does not impose any correlation between value added and student ability. The case of *Sorting on common ability* imposes a correlation of 1 between the value added of the Math teacher and the common ability factor θ_i, specifically by imposing a structure $\theta_i = a \times \theta_{j_M} + b$. Finally, the case of *Sorting on subject-specific ability* imposes a correlation of 1 between the value added of the Math teacher and the Math-specific ability factor $\theta_{i,M}$, specifically by imposing a structure $\theta_{i,M} = a \times \theta_{j_M} + b$.

E Descriptive statistics

Table E.5: Math teacher characteristics

	n	mean	sd	min	max	range	se
Experience	71,548	11.46	6.60	0.00	44.00	44.00	0.02
Age	71,548	39.99	8.16	22.00	67.00	45.00	0.03
At disadv. school	71,548	0.21	0.41	0.00	1.00	1.00	0.00
Perc. of time teaching	7,113	77.42	10.96	46.94	100.00	53.06	0.13
Pedag. grade (harm.)	15,022	46.42	4.17	29.00	60.00	31.00	0.03
Adv. qualif. (agrégé)	71,548	0.09	0.28	0.00	1.00	1.00	0.00
Gross monthly salary	70,122	2998.53	425.75	2047.37	4919.13	2871.76	1.61

Table E.6: French teacher characteristics

	n	mean	sd	min	max	range	se
Experience	77,510	12.18	6.53	0.00	39.00	39.00	0.02
Age	77,510	40.70	8.30	22.00	67.00	45.00	0.03
At disadv. school	77,510	0.21	0.41	0.00	1.00	1.00	0.00
Perc. of time teaching	9,297	76.21	11.19	50.00	100.00	50.00	0.12
Pedag. grade (harm.)	16,605	46.42	4.27	33.30	60.00	26.70	0.03
Adv. qualif. (agrégé)	77,510	0.08	0.28	0.00	1.00	1.00	0.00
Gross monthly salary	76,321	3021.24	435.77	2047.37	5200.80	3153.43	1.58

Table E.7: Student characteristics

	n	mean	sd	min	max	range	se
Age	2,450,480	14.32	0.57	12.00	18.00	6.00	0.00
Moved region since birth	2,450,480	0.33	0.47	0.00	1.00	1.00	0.00
Takes Ancient Greek	2,450,480	0.03	0.16	0.00	1.00	1.00	0.00
Takes Latin	2,450,480	0.16	0.36	0.00	1.00	1.00	0.00
Num. times taking exam	2,450,480	1.03	0.17	1.00	3.00	2.00	0.00
Passed 9th grade	2,450,480	0.76	0.43	0.00	1.00	1.00	0.00
French grade (pre-2016)	2,120,131	20.97	7.50	0.00	40.00	40.00	0.01
French grade (post-2016)	330,349	42.81	18.15	0.00	100.00	100.00	0.03
Math grade (pre-2016)	2,120,131	19.02	9.59	0.00	40.00	40.00	0.01
Math grade (post-2016)	330,349	47.75	24.73	0.00	100.00	100.00	0.04
At disadv. school	2,450,480	0.19	0.39	0.00	1.00	1.00	0.00
Scholarship student	2,450,480	0.23	0.42	0.00	1.00	1.00	0.00
Female	2,450,480	0.51	0.50	0.00	1.00	1.00	0.00
French national	2,450,480	0.97	0.18	0.00	1.00	1.00	0.00
Very high SES	2,450,480	0.27	0.44	0.00	1.00	1.00	0.00
High SES	2,450,480	0.19	0.39	0.00	1.00	1.00	0.00
Medium SES	2,450,480	0.33	0.47	0.00	1.00	1.00	0.00
Low SES	2,450,480	0.21	0.41	0.00	1.00	1.00	0.00
Parent employed	2,450,480	0.79	0.41	0.00	1.00	1.00	0.00
Parent retired	2,450,480	0.01	0.10	0.00	1.00	1.00	0.00
Parent unemployed	2,450,480	0.02	0.14	0.00	1.00	1.00	0.00
Parent other situation	2,450,480	0.05	0.21	0.00	1.00	1.00	0.00

Table E.8: Balance checks - schools in value-added analysis

	Coefficient	**S.E.**
Average teacher age	-1.1103***	0.0879
Average class size	0.1946***	0.0442
Average cohort size	-18.9017***	0.7764
Average teacher experience	1.2675***	0.0664
Average french grade	0.0179*	0.0059
Average math grade	0.0135	0.0072
Average number of teachers per year	-0.8457***	0.0375
Average number of french teachers per year	-0.4592***	0.0209
Average number of math teachers per year	-0.4009***	0.0187
Average DNB pass rate	0.0049*	0.0017
Average pedagogical grade	-0.0196	0.0113
Average perc. pedagogical grade	-0.0074 **	0.0035
Average perc. agregation	0.0077***	0.0021
Average perc. CAPES	0.0492***	0.0022
Perc. French students	0.0053***	0.0010
Perc. male students	3e-04	0.0006
Average perc. male teachers	0.0126***	0.0034
Perc. scholarship stusdents	-0.0133***	0.0030
Perc. students SES 1	0.0106***	0.0031
Perc. students SES 2	0.0061***	0.0012
Perc. students SES 3	0.0034	0.0018
Perc. students SES 4	-0.0033	0.0030
Perc. students taking Ancient Greek	3e-04	0.0010
Perc. students taking Latin	3e-04	0.0016
Average teacher salary	-73.887***	4.5570
Average number of times taking DNB exams	6e-04	0.0004
Disadvantaged school	-0.019 **	0.0088

Note: The table presents the OLS regression coefficients of a regression which compares the treated schools' characteristics to those of all schools (i.e. including the treated).
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table E.9: Network statistics

	Mean	**S.d.**
Math teachers		
Number of networks per teacher within school	4.42	1.79
Number of times a network is observed	5.6	6.54
Number of students per network	135	158.7
Number of French link teachers per Math teacher	2.75	1.46
Number of unique networks	39,094	
French teachers		
Number of networks per teacher within school	5.36	2.09
Number of times a network is observed	4.85	5.78
Number of students per network	117	139.9
Number of Math link teachers per French teacher	2.4	1.26
Number of unique networks	58,232	

F Results

F.1 First stage

The coefficients of the first stage are shown in Table F.10.[31]

It is worth noting that all factors $\mathbf{X}_i$ have a significantly different impact on student grades depending on whether we examine Math or French grades, indicating that certain factors are more important for the performance of a student in a certain subject. While coefficients across subjects are nevertheless generally in the same ballpark, it is interesting to note that male students perform slightly better in Math (0.03 s.d.), and significantly worse in French (-0.38 s.d.).

In addition, the socio-economic status of students (expressed both by their scholarship status and their parents' SES) is particularly important for the DNB grades of the students. Compared to a student with a very high SES status, the Math (French) grade of a low SES student is 0.62 (0.58) s.d. lower on average, for two students who study at the same school in the same year, and are thus exposed to a very similar school environment.

Taking Ancient Greek or Latin are also very good indicators of the performance of a student, which likely has to do with the SES of a student, due to the fact that the take-up of the two languages is often seen as a mechanism by more strategic parents of sorting their children into better classrooms.

Despite the number of student-level controls and the specific focus on the within-school variation in estimates, these observable characteristics seem to only explain about 30% (29%) of the variation in Math (French) DNB grades.

[31] Note that almost all controls are typical for the value-added literature, with the exception of the dummies " Taking Ancient Greek" and " Taking Latin". The latter two are included due to anecdotal evidence that parents use the two optional subjects in order to sort their children into classrooms of higher performance/ability.

Table F.10: First stage of value-added estimation

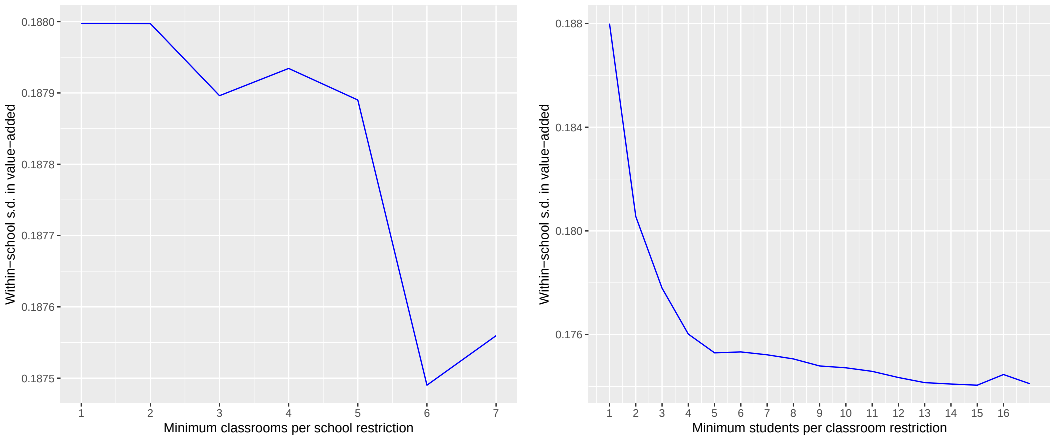
	Math grade	French grade
Male	0.030***	-0.379***
	(0.001)	(0.001)
Age	-0.364***	-0.333***
	(0.001)	(0.001)
Scholarship student	-0.141***	-0.114***
	(0.001)	(0.001)
High SES	-0.267***	-0.224***
	(0.002)	(0.002)
Medium SES	-0.404***	-0.351***
	(0.001)	(0.001)
Low SES	-0.618***	-0.582***
	(0.002)	(0.002)
Moved department	0.018***	0.037***
	(0.001)	(0.001)
Taking Ancient Greek	0.411***	0.399***
	(0.003)	(0.003)
Taking Latin	0.540***	0.552***
	(0.001)	(0.001)
School FE	Yes	Yes
Teacher experience FE	Yes	Yes
Year FE	Yes	Yes
Nationality FE	Yes	Yes
Main guardian type FE	Yes	Yes
Observations	2,807,201	2,807,201
R^2	0.306	0.289

Note: The table presents the OLS regression results for the association of observable student characteristics with student grades in Math and French. Student grades are standardised by subject and year. The base SES category is that of "Very high SES". The dummy "Moved department" is equal to 1 if a student's birth department is different from his current department of living.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

F.2 Robustness

Figure F.7: Average within-school s.d. in value-added given different restrictions on the minimum number of classrooms per school (LHS) and the number of students per classroom (RHS)

(a) Minimum number of classrooms (b) Minimum number of students



F.3 Explaining the value added estimates conditionally

	Math VA	French VA
Age	0.099*	−0.014
	(0.055)	(0.038)
Age^2	−0.001**	0.00000
	(0.001)	(0.0004)
Experience	−0.017	−0.003
	(0.023)	(0.018)
$Experience^2$	0.0002	0.0002
	(0.001)	(0.001)
Total num. students	0.002***	−0.0003
	(0.0005)	(0.0004)
Male	−0.184**	−0.242***
	(0.085)	(0.074)
Pedag. grade (harm.)	0.432***	0.220**
	(0.167)	(0.104)
Adv. qualif. (agrégé)	−0.309	0.072
	(0.226)	(0.144)
Hours teaching	−0.306**	0.062
	(0.125)	(0.084)
Gross monthly salary	0.001***	0.0002
	(0.0002)	(0.0001)
Mover	−0.031	0.018
	(0.093)	(0.064)
Leaver	0.081	−0.020
	(0.154)	(0.110)
School FE	Yes	Yes
Observations	4,194	5,010
R^2	0.613	0.590

Note: The table presents the OLS regression results for the association of observable teacher characteristics with the estimated teacher value added. The estimates are available for a small subsample of teachers due to missing data in the characteristics. The correlations are conditional on school fixed effects to reflect the fact teacher value added estimates are comparable only within school. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$