Social Spillovers in the Classroom: Identification, Estimation, and Policy Analysis

Santiago Pereda-Fernández*

Banca d'Italia

November 7, 2016

Abstract

I present a method to jointly estimate social spillovers in the classroom and the distributions of teacher and student effects. This method builds on Graham (2008) and is based on the covariance and higher order moments restrictions of the test scores and requires the random assignment of teachers and students to classrooms. Using the Tennessee Project STAR dataset, I find sizable spillovers in kindergarten classrooms and departures from normality of the teacher and student ability distributions. I also find that reducing class size has a positive effect on mean performance but it increases the inequality. Based on these estimates, I perform several input-neutral policy counterfactuals involving teachers and students assignment rules, and changing the distribution of class sizes. For the latter, I derive an optimal class size distribution rule, which increases mean test scores and reduces the overall variance.

Keywords: Academic Performance, Class Size, Counterfactuals, Higher-Order Moments, Peer Effects

JEL classification: C31, C51, I21, I28

*Banca d'Italia, Via Nazionale 91, 00184 Roma, Italy. This is a revised version of the first chapter of my dissertation. I would like to thank my supervisor, Bryan Graham, for his continuous support and advice with this project. I would also like to thank Stéphane Bonhomme for his supervision during the year I spent at CEMFI. I am also grateful to Manuel Arellano, Samuel Bentolila, Guillermo Caruana, Hilary Hoynes, Patrick Kline, Mónica Martínez Bravo, Filip Matějka, Pedro Mira, James Powell, Enrique Sentana, Jesse Rothstein, Frank Vella, Paolo Zacchia, two anonymous referees, and seminar participants at Banca d'Italia, CEMFI, CERGE-EI, EIEF, University of California, Berkeley and University of Edinburgh for their helpful comments and suggestions. The views expressed in this paper are my own and do not necessarily reflect those of the Banca d'Italia. All remaining errors are my own. I can be reached via email at santiago.pereda@bancaditalia.it

1 Introduction

It is an empirical fact that there are persistent differences in mean test scores across classes (Hanushek, 1971). Some of these differences can be accounted for by observed characteristics, such as class size. However, they can also be due to unobservable factors, such as teacher quality, or the ability of a student's peers. There are many studies that provide a partial analysis of the effect of each of these factors in isolation. For instance, Angrist and Lavy (1999) found that reducing predicted class size by ten students would increase the average class test scores by about 0.25 standard deviations. Despite the inherent interest of such analyses, principals or policy makers face a different problem: given a fixed number of students and teachers, how should they be allocated? What should be the size of each class? Further, if the policy maker cares about inequality, it is not sufficient to focus on the average performance, and instead one should pay attention to the distributional effects of a policy.[1]

In this paper I make the following contributions: first, I propose a method to jointly identify and estimate social spillovers in the classroom, and the distributions of teacher and student effects. This method, which extends Graham (2008), is based on the covariance and higher order moments restrictions in students' test scores. Further, by allowing teacher and student effects to be heteroskedastic in class size, I can assess the distributional effect of class size on test scores. Second, I apply this method to the Tennessee Project STAR experiment dataset, finding substantial spillovers at the kindergarten level, as well as distributions of the teacher and student effects that are not normally distributed. Moreover, I find that reducing class size has a positive effect on mean test scores, but it increases their dispersion. Third, using these estimates I carry out several counterfactual experiments involving reassignment of teachers and students into classrooms, and changing the distribution of class sizes, obtaining different distributions of test scores. Some of these counterfactuals simultaneously increase mean test scores and reduce their dispersion, without varying the educational inputs.

A major challenge in this setup is to disentangle between the teacher effects and the student peer effects. If a classroom is assigned a teacher that is more effective than the

current one, then their students' test scores will increase. However, if there are spillovers among students, it will lead to a further endogenous change in students' test scores. This problem was already noted by Manski (1993), who referred to it as the *reflection problem*.

In the literature of peer effects estimation, most identification methods rely on the partial overlap of the reference groups (Bramoullé et al., 2009), or on some external variation that can be used as an instrument (Angrist, 2014).[2] In this paper, I consider a different framework in which the data has group (class) structure, and therefore the reference groups are homogeneous. In contrast with other studies (Hoxby, 2000; Arcidiacono et al., 2012; Burke and Sass, 2013), it is not required to have a panel, and a single observation per student suffices. In this setup, I extend Graham (2008) model and I propose a minimum distance estimator that combines moments of different order and accommodates missing test scores at random in a simple way. This method requires the double randomization of students and teachers into classrooms, and class size variation, and it allows to jointly estimate the social multiplier and the distributions of teacher and student effects. The social multiplier, which measures the size of the spillovers, is linked to a structural model of peer effects, and it reflects both endogenous and exogenous effects in Manski's taxonomy.

Many policies result in a change in the distribution of the students' performance that is not accounted for by the mean effect. For example, tracking of students by ability would reduce within class inequality at the cost of increasing between class inequality, but the overall effect is not clear. Similarly, many studies focused on the mean effect of class size on test scores, but not on its distributional effects. In this paper I take an approach based on the estimation of the higher order moments of the distributions of teacher and student latent effects. These distributions are interesting by themselves, since at least teacher effects are often assumed to be normally distributed, and a departure from this assumption is likely to have implications on the evaluation of teachers' performance.[3]

I use the Tennessee project STAR dataset, which satisfies the assumptions required for the identification of the peer effects. This experiment was originally designed to estimate

the impact of class size on test scores at early childhood education. During kindergarten, students and teachers were randomly assigned to classrooms of different size, and they were tracked during the subsequent years. Recent studies have found test scores at this early stage to be correlated with economic outcomes later in life (Chetty et al., 2011). Consequently, an intervention during kindergarten has the potential of affecting lifetime earnings and even a high rate of return (Heckman et al., 2010).

The results show the existence of strong spillovers: increasing the average ability of the classmates of a student by one standard deviation results in a mean increase of test scores of around 0.45 standard deviations. Moreover, teachers also have a large impact on students' performance, and being assigned a teacher one standard deviation more efficient would result in an increase of test scores by 0.11 to 0.15 standard deviations. The results indicate that the distributions of teacher quality and student ability depart from the usual normality assumption, showing different degrees of skewness and kurtosis. Moreover, the student effects distribution is heteroskedastic in class size: the larger the class, the smaller variance of the student effect. The dispersion of the teacher effect is constant for different values of class size, despite its mean effect being negatively correlated with class size. Hence, increasing class size reduces the within class inequality, measured as the variance of the test scores, at the cost of reducing the mean performance.

Based on these estimates I conduct several counterfactual policy experiments. Given a number of teachers and enrolled students, a principal can choose how many students to allocate in each class, and who to assign to each of them, as well as who would be their teacher. The results show that assignment of good teachers to large classrooms increases mean test scores and reduces inequality. Tracking of students increases the test scores of high achievers at the cost of reducing the performance of low achievers. On the other hand, detracking has a small effect on mean test scores, while at the same time it decreases the level of inequality. Finally, I also consider the problem of choosing the optimal class size distribution, which not only increases mean performance, but also reduces inequality.

The rest of the paper is organized as follows: section 2 discusses the identification of the peer effects together with the distributions of teacher and student effects, and the estimator based on this results is proposed in section 3; section 4 describes the Tennessee Project STAR experiment dataset and presents the estimation results; using these estimates I run several counterfactuals in section 5, and section 6 concludes.

2 Identification

Denote the test score of student i in classroom c by y_{ic}, the teacher effect by α_c, and the student effects by $\{\varepsilon_{ic}\}_{i=1}^{N_c}$. These effects are latent variables that reflect factors such as ability, that are unobservable. Let the test scores be given by the linear-in-means equation

$$y_{ic} = \alpha_c + (\gamma - 1)\,\overline{\varepsilon}_c + \varepsilon_{ic} \tag{1}$$

where γ is the social multiplier that captures the strength of the peer effects, which can be endogenously derived from a model of social interactions in the classroom.[4] In the presence of peer effects ($\gamma > 1$) class composition has an impact on individual test scores, and improving the composition of a student's peers would enhance his test scores.

I consider a network with a group structure, *i.e.* all students in a classroom form a group and they affect each other, so the first moment is not sufficient to distinguish between the teacher and peer effects.[5] Hence, I normalize student effects to have conditional zero mean, such that the conditional expectation of test scores equals the conditional expectation of teacher effect, *i.e.* $\mathbb{E}\left[y_{ic}|N_c\right] = \mathbb{E}\left[\alpha_c|N_c\right]$.

The identification strategy combines Graham (2008), who identifies the social multiplier by comparing the differences between two types of classes of the between and the within class variances of the test scores, with the linear independent factor framework in Bonhomme and Robin (2009).[6] With respect to the latter, there are two main differences: vectors can be of different dimension and there are missing observations. These issues are overcome because of the double random assignment.

Rather than identifying the social multiplier in isolation, I also identify the distributions of teacher and student effects, which are necessary to conduct the counterfactual analysis. Moreover, I relax the assumption that teacher effects are homoskedastic in class size, and allow both teacher and student effects to display different types of heteroskedasticity. Also, with respect to Graham (2008) approach, the treatment of missing test scores is simpler and more convenient to implement, particularly for higher order moments.[7]

By equation 1, the covariances of the test scores depend on the variances of all the individual factors, as well as the covariances among them. If there is sorting of students or teachers, then the covariances reflect both the spillovers and the cross correlation in agents' abilities. The following assumptions rule out the existence of sorting:

Assumption 1. *Conditional double randomization:* $\left(\alpha_c, \{\varepsilon_{ic}\}_{i=1}^{N_c}\right)$ *are jointly independent given* N_c.

Assumption 2. *Class size is independent of students and teacher's assignment mechanism.*

Denote by Y_c the vector with the demeaned test scores, and by $X_c \equiv (\alpha_c, \varepsilon_{1c}, ..., \varepsilon_{N_c})'$, the vector with the teacher and student effects in class c. Then,

$$Y_c = \Lambda(\gamma; N_c) X_c \tag{2}$$

where $\Lambda(\gamma; N_c) \equiv \left(\iota_{N_c}, I_{N_c} + \frac{\gamma-1}{N_c}\iota_{N_c}\iota'_{N_c}\right)$, I_N is the identity matrix of dimension N, and ι_N is a vector of ones of dimension N. Each of the rows of the matrix $\Lambda(\gamma; N_c)$ contains the contribution of the teacher and student effects to the test score of a single student.[8] Conditional on class size, the variance matrix of the test scores, Σ_{Y,N_c}, depends on two elements: the variance of the teacher effect, $\sigma^2_\alpha(N_c)$, and the variance of the student effect, $\sigma^2_\varepsilon(N_c)$. Under assumptions 1 and 2, the covariance of the test scores of students i and j is given by

$$Cov(\tilde{y}_{ic}, \tilde{y}_{jc}|N_c) = \sigma^2_\alpha(N_c) + \left[\frac{\gamma^2-1}{N_c} + \mathbf{1}(i=j)\right]\sigma^2_\varepsilon(N_c) \tag{3}$$

The same strategy can be applied to higher order moments. If the teacher and student effects are not normally distributed, the higher order cumulants provide more features of

their distributions, along with overidentifying restrictions for the social multiplier. In this paper I work with cumulants, which also characterize the distribution of a random variable.[9] Assumption 1 and the linear factor representation of test scores make working with cumulants very tractable: the Rth cumulant of the test scores is a weighted sum of the Rth cumulants of the teacher and student effects. In particular, the third and fourth cumulants of the students' tests scores are given by

$$\kappa_3\left(\tilde{y}_{ic},\tilde{y}_{jc},\tilde{y}_{hc}|N_c\right)=\kappa_3\left(\alpha_c|N_c\right)+\left\{\frac{\left(\gamma-1\right)^2\left(\gamma-2\right)}{N_c^2}\right.$$
$$\left.+\frac{\gamma-1}{N_c}\left[\mathbf{1}\left(i=j\right)+\mathbf{1}\left(i=h\right)+\mathbf{1}\left(j=h\right)\right]+\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=h\right)\right\}\kappa_3\left(\varepsilon_{ic}|N_c\right)$$
$$\kappa_4\left(\tilde{y}_{ic},\tilde{y}_{jc},\tilde{y}_{hc},\tilde{y}_{kc}|N_c\right)=\kappa_4\left(\alpha_c|N_c\right)+\left\{\frac{\left(\gamma-1\right)^3\left(\gamma-3\right)}{N_c^3}\right.$$
$$+\frac{\left(\gamma-1\right)^2}{N_c^2}\left[\mathbf{1}\left(i=j\right)+\mathbf{1}\left(i=h\right)+\mathbf{1}\left(i=k\right)+\mathbf{1}\left(j=h\right)+\mathbf{1}\left(j=k\right)+\mathbf{1}\left(h=k\right)\right]$$
$$+\frac{\gamma-1}{N_c}\left[\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=h\right)+\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=k\right)\mathbf{1}\left(i=h\right)\mathbf{1}\left(i=k\right)+\mathbf{1}\left(j=h\right)\mathbf{1}\left(j=k\right)\right]$$
$$\left.+\mathbf{1}\left(i=j\right)\mathbf{1}\left(i=h\right)\mathbf{1}\left(i=k\right)\right\}\kappa_4\left(\varepsilon_{ic}|N_c\right)$$

The variance has two different permutations, either $i=j$ or $i\neq j$, but the third and fourth order cumulants have more.[10] Consequently, the cross cumulants of the test scores depend differently on these permutations. To avoid working with arrays of different order, I apply the operator *vech* to the second to fourth order arrays of the cumulants of the test scores, and express the resulting vectors with the non repeated elements of these arrays as a linear function in the cumulants of the teacher and student effects:[11]

$$\omega_{Y,N_c}^2\equiv vech\left(\Sigma_{Y,N_c}\right)=\Lambda_2\left(\gamma;N_c\right)D_2\left(\alpha_c,\varepsilon_{ic}|N_c\right)$$
$$\omega_{Y,N_c}^3\equiv vech\left(\Gamma_{Y,N_c}\right)=\Lambda_3\left(\gamma;N_c\right)D_3\left(\alpha_c,\varepsilon_{ic}|N_c\right)$$
$$\omega_{Y,N_c}^4\equiv vech\left(\Omega_{Y,N_c}\right)=\Lambda_4\left(\gamma;N_c\right)D_4\left(\alpha_c,\varepsilon_{ic}|N_c\right)$$

where $D_r\left(\alpha_c,\varepsilon_{ic}|N_c\right)\equiv\left(\kappa_r\left(\alpha_c|N_c\right),\kappa_r\left(\varepsilon_{ic}|N_c\right)\right)'$ for $r=2,3,4$, and the $\Lambda_r\left(\gamma;N_c\right)$ matrices are defined in appendix F.

The last ingredient required for the identification is variation in class size. Equation 1 does not explicitly model the effect of class size on teacher and student effects. In principle,

some students are relatively more efficient in small classrooms than other students, and similarly some teachers are more adept at teaching in small classrooms. Hence, the variance and higher order moments of these effects can vary with class size. In this paper I consider three different models: homoskedastic effects and two different types of heteroskedasticity.

In the homoskedastic effects model, the variances of the teacher and student effects are constant with respect to class size. The second model is heteroskedastic in class type, with classes being either small or large depending on the number of students. In this model, each agent has a different potential outcome in each of the class types. Mathematically, $\alpha_c = \alpha_{0c}\mathbf{1}(small) + \alpha_{1c}\mathbf{1}(large)$ and $\varepsilon_{ic} = \varepsilon_{0ic}\mathbf{1}(small) + \varepsilon_{1ic}\mathbf{1}(large)$. The third model is a random coefficients model in class size, *i.e.* $\alpha_c = \alpha_{0c} + \alpha_{1c}N_c$ and $\varepsilon_{ic} = \varepsilon_{0ic} + \varepsilon_{1ic}N_c$, and the variance is a polynomial of order two of class size.[12]

The cumulants of the teacher and student effects are either constant (homoskedasticity), different for small and large classrooms (class type heteroskedasticity), or a polynomial of order R of class size (random coefficients), *i.e.* $\kappa_R(\alpha_c|N_c) = \sum_{r=0}^{R}\mu_{\alpha,R,r}N_c^r$, and $\kappa_R(\varepsilon_c|N_c) = \sum_{r=0}^{R}\mu_{\varepsilon,R,r}N_c^r$. The number of parameters in this system of equations equals to 7, 13, and 25, respectively. Given H distinct class sizes, the total number of moment restrictions is $10H$.[13] The total number of parameters for each of the models equals 3, 5, and 7.[14] Hence, if there are classes of at least four different sizes, it is possible to identify all the parameters in all three models using just the variances and covariances.

3 Estimation

The first step is to estimate equation 1 with OLS, and denote the residuals by $\hat{y}_{ic}$. For class c, define the vectors $\hat{\omega}^2_{Y,c}$, $\hat{\omega}^3_{Y,c}$ and $\hat{\omega}^4_{Y,c}$ as the vectors resulting from applying the *vech* operator to $\hat{\Sigma}_{Y,c}$, $\hat{\Gamma}_{Y,c}$, and $\hat{\Omega}_{Y,c}$, which are the arrays of dimension 2, 3, and 4 respectively, with generic elements

$$\hat{\Sigma}_{Y,c}(i,j) = \hat{y}_{ic}\hat{y}_{jc}$$

$$\hat{\Gamma}_{Y,c}(i,j,h) = \hat{y}_{ic}\hat{y}_{jc}\hat{y}_{hc}$$

$$\hat{\Omega}_{Y,c}(i,j,h,k) = \hat{y}_{ic}\hat{y}_{jc}\hat{y}_{hc}\hat{y}_{kc} - \left[\hat{\sigma}_Y^2(i,j)\,\hat{\sigma}_Y^2(h,k) + \hat{\sigma}_Y^2(i,h)\,\hat{\sigma}_Y^2(j,k) + \hat{\sigma}_Y^2(i,k)\,\hat{\sigma}_Y^2(j,h)\right]$$

where the estimator of the covariance term between students l and m is given by

$$\hat{\sigma}_Y^2(l,m) = \frac{\Sigma_{c=1}^{C}\Sigma_{i=1}^{N_c}\hat{y}_{ic}^2}{\Sigma_{c=1}^{C}N_c}\mathbf{1}(l=m) + \frac{\sum_{c=1}^{C}\sum_{i=1}^{N_c-1}\sum_{j=i+1}^{N_c}\hat{y}_{ic}\hat{y}_{jc}}{\frac{1}{2}\sum_{c=1}^{C}N_c(N_c-1)}\mathbf{1}(l\neq m)$$

In words, the $\hat{\omega}_{Y,c}^{j}$ vectors contain the sample analogues of the cumulants of all possible combinations of j test scores with repetition but without ordering them. For the variance, it would include all of the N_c individual variances and the $\frac{N_c(N_c-1)}{2}$ distinct covariances, and similarly for higher order cumulants. These vectors are concatenated, creating the vector $\hat{\omega}_Y$. Similarly, the Λ_{j,N_c} and D_j matrices are suitably concatenated to create the matrices Λ and D. Given the weighting matrix W_C, the minimum distance estimator is the solution to:

$$\hat{\theta}_{MD} = \arg\min_{\theta}\left(\hat{\omega}_Y - \Lambda D\right)' W_C\left(\hat{\omega}_Y - \Lambda D\right) \tag{4}$$

where $\theta \equiv \left[\gamma, \kappa_2(\alpha_c), \kappa_3(\alpha_c), \kappa_4(\alpha_c), \kappa_2(\varepsilon_{ic}), \kappa_3(\varepsilon_{ic}), \kappa_4(\varepsilon_{ic})\right]'$ under the assumption of homoskedastic teacher and student effects. If the effects are heteroskedastic, the vector θ is appropriately defined.[15]

There are two compelling reasons not to use the identity matrix: the higher the order of the cumulant, the noisier it is, and the higher the weight it receives in the estimation.[16] To address the first problem, I follow Cragg (1997), who respectively gives weights $\frac{1}{2}$, $\frac{1}{15}$ and $\frac{1}{96}$, to second, third and fourth order cumulants.[17] The second problem is overcome by weighting each moment by the inverse of the number of cumulants of the same order, *i.e.* $\binom{N_c+R-1}{R}^{-1}$. Standard errors are calculated using the robust White formula with clusters at the school level.

4 Estimation of Peer Effects in the Kindergarten

4.1 Data

I use the data from the Tennessee Project STAR experiment, whose original goal was to estimate the impact that a class size reduction policy would have on students achievement.[18]

This experiment is also well suited for the analysis of spillovers in the classroom, and it has previously been used to estimate peer effects (Graham, 2008; Chetty et al., 2011). Classes were split into three types: small, regular, and regular with aide.[19] Small classes had between 13 and 17 students, and the other two types of classes would have between 22 and 25 students each, with the difference that regular with aide classes had a full time teacher's aide, and regular classes did not. In order to be eligible for participation, the number of students enrolled in each school had to be high enough to accommodate at least one class of each type. Once class sizes were determined, students were randomly sorted into class type, and teachers were randomly matched into class type.[20]

The data consists of 6308 kindergarten students distributed across 325 classrooms.[21] At the end of the academic year, students took the *Stanford Achievement Tests in Mathematics and Reading*. No measure of ability or pretreatment test scores are available. To make the results comparable with other studies, test scores are normalized to have mean zero and variance one. Among those students who were enrolled, test scores are observed for the majority of the students, but not all of them.[22] Table 1 shows the absolute frequency of class size, which ranges from 11 to 28 students, and values between 13 and 17, and between 21 and 24 exhibit the highest frequencies. As a result, the between and within variances are much more precise for these values of class size.

4.2 First Moment Estimates

Table 2 summarizes the results of the regression of the equation in levels.[23] Given that the randomization took place within schools, I control for differences across schools by including school dummies in all specifications. The class size coefficient is negative in every specification and it is significant at the 99% confidence level, both for the mathematics and reading test scores. Classes of regular size with aide have a negative coefficient associated to them, although this may be because this variable is correlated with large sized classes.

Table 1: Class size distribution and variance decomposition

Class size	11	12	13	14	15	16	17	18	19
Frequency	3	5	19	23	24	31	29	3	13
$Mean_M$	-0.03	0.08	0.06	0.45	-0.14	0.14	0.08	-0.25	0.09
$Var_{B,M}$	0.04	0.15	0.28	0.31	0.44	0.40	0.25	0.06	0.28
$Var_{W,M}$	0.65	0.73	0.66	0.73	0.78	0.75	0.69	0.65	0.63
$Mean_R$	-0.32	0.20	0.17	0.39	-0.07	0.14	0.10	-0.39	-0.02
$Var_{B,R}$	0.11	0.53	0.26	0.33	0.37	0.32	0.33	0.01	0.47
$Var_{W,R}$	0.40	1.03	0.67	0.62	0.70	0.88	0.64	0.28	0.69
Class size	20	21	22	23	24	25	26	27	28
Frequency	14	27	40	36	32	12	6	7	1
$Mean_M$	-0.16	-0.15	0.13	0.09	-0.15	-0.09	-0.67	0.01	0.45
$Var_{B,M}$	0.14	0.22	0.24	0.53	0.25	0.32	0.44	0.28	0
$Var_{W,M}$	0.53	0.65	0.68	0.70	0.66	0.64	0.50	0.53	0.45
$Mean_R$	-0.11	-0.06	0.14	-0.03	-0.19	0.01	-0.69	0.00	0.09
$Var_{B,R}$	0.17	0.29	0.33	0.30	0.27	0.28	0.30	0.27	0
$Var_{W,R}$	0.86	0.70	0.67	0.69	0.67	0.60	0.48	0.53	0.29

Notes: The subscripts B, W, M, and R respectively denote between variance, within variance, mathematics test, and reading test.

Table 2: OLS estimates of the equation in levels

	Mathematics		Reading	
	(1)	(2)	(1)	(2)
Class Size	-0.022***	-0.021***	-0.023***	-0.020***
	(0.003)	(0.004)	(0.003)	(0.004)
Regular with aide	-	-0.026	-	-0.053**
		(0.027)		(0.027)
School dummies	✓	✓	✓	✓

Notes: Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels.

4.3 Variance and Higher Order Cumulants Estimates

I consider three specifications, in which student effects are respectively homoskedastic, heteroskedastic in class type, and a random coefficients model in class size, whereas teacher effects are considered homoskedastic in all three specifications[24] For each of them there are three sets of estimates, one which uses only the variances, another one that uses also the third order cumulants, and another one that uses up to the fourth order cumulants.[25]

Table 3 shows the estimates of the social multiplier, the standard deviation, and the third and the fourth cumulants of the teacher effect.[26] For the mathematics exam, the social multiplier is between 1.4 and 1.8, substantially larger than one, though only barely significant.[27] When the student effects distribution is assumed to be homoskedastic, the

Table 3: Variance and higher order teacher effect cumulants estimates

	Mathematics Test Scores								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$-1	0.854**	0.868**	0.867**	0.545*	0.564*	0.564*	0.520*	0.544*	0.544*
	(0.374)	(0.395)	(0.374)	(0.299)	(0.299)	(0.300)	(0.311)	(0.311)	(0.312)
$\hat{\sigma}_\alpha$	-	-	-	0.156	0.149	0.149	0.164	0.156	0.156
				(0.109)	(0.115)	(0.116)	(0.106)	(0.113)	(0.113)
$\hat{\kappa}_3(\alpha_c)$	-	0.007	0.007	-	0.008	0.008	-	0.008	0.008
		(0.012)	(0.010)		(0.010)	(0.010)		(0.010)	(0.010)
$\hat{\kappa}_4(\alpha_c)$	-	-	-0.076***	-	-	-0.075***	-	-	-0.076***
			(0.009)			(0.010)			(0.010)
	Reading Test Scores								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$-1	0.791*	0.776*	0.733*	0.553	0.545	0.505	0.466	0.471	0.427
	(0.413)	(0.456)	(0.416)	(0.349)	(0.341)	(0.344)	(0.371)	(0.361)	(0.364)
$\hat{\sigma}_\alpha$	-	-	-	0.091	0.095	0.116	0.132	0.130	0.147
				(0.222)	(0.203)	(0.163)	(0.152)	(0.149)	(0.129)
$\hat{\kappa}_3(\alpha_c)$	-	0.001	0.002	-	0.004	0.004	-	0.004	0.005
		(0.014)	(0.011)		(0.011)	(0.011)		(0.010)	(0.011)
$\hat{\kappa}_4(\alpha_c)$	-	-	-0.072***	-	-	-0.070***	-	-	-0.069***
			(0.012)			(0.012)			(0.012)

Notes: Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels. Specifications 1 to 3 assume that moments of student effects are the same for all students (*i.e.*, homoskedastic effects); specifications 4 to 6 allow them to have different values for students in small and large classes; specifications 7 to 9 assume that student effect is a random coefficient in class size, and thus their cumulants are polynomials in class size.

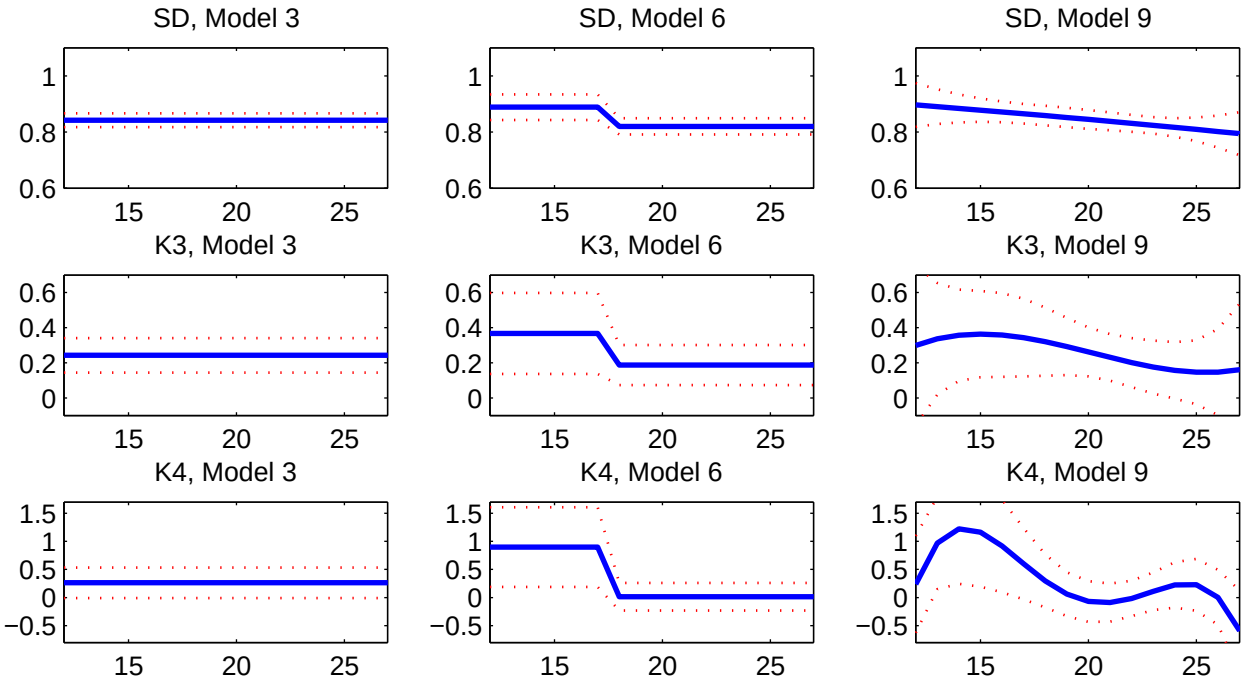
estimated variance of teacher's quality is negative in all cases, but these estimates are not significant.[28] If student effects are heteroskedastic, then they become positive and close to 0.15, in line with previous results found in the literature (Hanushek and Rivkin, 2010). The estimates of the third cumulant of the teacher effect are slightly positive but insignificant in all specifications, which suggests that the distribution is roughly symmetric. The estimates of the fourth cumulant are negative and significant, implying that the distribution of the teacher effect has thinner tails than the normal distribution (playtkurtic). Thus, even though there is variability in teacher quality, there are very few extremely effective or ineffective teachers.

Figure 1 shows the estimates of the standard deviation, the third and the fourth cumulants of the student effect for specifications 3, 6, and 9, for the mathematics test scores. The estimates of the standard deviation of the student effects range between 0.8 and 0.9, and when it is allowed to be heteroskedastic, there is a decreasing pattern as class size increases. To get an idea of the magnitude of the spillovers operating through the social multiplier, if I changed the classmates of a student, and the ability of the new classmates was on average one standard deviation larger, it would increase his test score by around 0.45 standard deviations. This number is very close to the one found by Carman and Zhang (2012) for kindergarten students in China (0.4).

The third cumulant is positive and significant, and similarly to the variance, it varies across different class sizes. The estimates are larger for smaller classes, which means that the smaller the class size, the more asymmetric the distribution is. In contrast with the distribution of teacher effects, the fourth cumulant is either positive (small classrooms) or significantly equal to zero (large classrooms). Hence, student's quality is more dispersed, and there is a substantial proportion of particularly good and bad students, and the performance is more heterogeneous in small classrooms. Finally, the precision of the estimates of the social multiplier is almost unaffected by the inclusion of the higher order cumulants in the estimation.

The results for the reading tests are qualitatively similar, but less precisely estimated.

Figure 1: Estimates of the standard deviation, third and fourth cumulants of student effect, mathematics test scores



Notes: The dotted line represents the 95% confidence interval. Standard errors computed for each class size using the delta method.

The most noticeable differences are the third and fourth cumulants of the distribution of student effects, which are larger than those found for the mathematics exam. In terms of the efficiency gained by using more cumulants in the estimation, the results are better than for the mathematics tests: the estimates of the social multiplier are approximately 2% more precise, and the standard error of the standard deviation of teacher effect gets significantly smaller by including the third and fourth cumulant, with gains of about 25% and 15% in each of the two heteroskedasticity models.

Equation 1 is one of the possible ways to model social interactions in the classroom. Since this model is not necessarily correctly specified, one would like to test whether the estimated peer effects are driven by the assumed functional form or if they reflect actual spillovers. I address this concern in appendix G, where I propose a test statistic that allows the social spillovers to be incorrectly specified, and whose distribution is known under the null hypothesis of no spillovers, *i.e.* $\gamma = 1$. For almost every specification, the test fails to

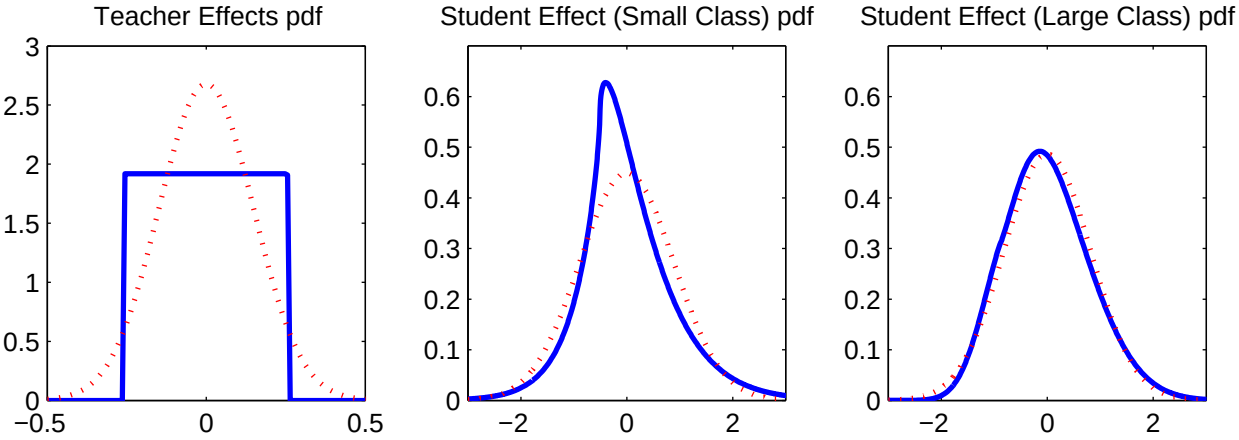
accept the hypothesis of no spillovers.

4.4 Non-Normally Distributed Teacher and Student Effects

The results show that the third and fourth order cumulants of student effects are significantly different from zero, and thus non-normal. It is a common practice to assume normality when the distribution of the unobservables is unknown, so I compare the normal distribution to the Skew Exponential Power $(\mu, \sigma, \lambda, \alpha)$ distribution: the second to fourth cumulants of the estimated distributions of teacher and student effects in specification 6 are fitted to the SEP distribution, and then compared to the normal distribution that has the same variance.[29]

Figure 2 shows the pdf of the teacher and student effects under normality and when the effects follow an unrestricted SEP distribution. The differences between the two distributions are quite marked for the teacher effects and the student effects in small classes, but not so much for student effects in the large ones. Interestingly, the distribution of the teacher effects has such a small fourth moment, that its support is a closed interval.

Figure 2: SEP and normal distributions for student and teacher effects



Notes: The solid lines represent the estimated SEP pdf, and the dotted lines represent the estimated normal pdf. The cumulants second to fourth of the SEP distribution have been fitted to those estimated in model 6 for a class with 15 students. For the normal distribution only the variance was fitted.

5 Counterfactuals and Policy Analysis

I consider two different types of input-neutral counterfactuals, those in which students and teachers are sorted based on their ability, and those in which the class size distribution is changed.[30] Input-neutral counterfactuals isolate the effects coming from an increase or a reduction in the resources at disposal. This is an important factor when one wants to evaluate the effect of reducing the average class size, which requires hiring new teachers, who may not be as effective as the old ones, and would therefore offset some of the potential gains from such policy.

I run a Monte Carlo (1000 repetitions) in which I draw teacher and student effects from the SEP distribution with the parameters implied by the estimates from specification 6, and then evaluate the distribution of test scores resulting from each counterfactual.[31] In all cases, the changes in the distribution are with respect to the case in which students and teachers are randomly assigned and the distribution of class sizes is the one observed in the data.[32]

5.1 Changing the Teacher and Student Assignment Rules

Consider the following four counterfactuals: (1) matching best teachers to largest classrooms, random assignment of students; (2) random assignment of teachers, tracking of students, best students assigned smallest classrooms; (3) matching best teachers to largest classrooms, tracking of students, best students assigned smallest classrooms; (4) random assignment of teachers, detracking of students, random assignment into classrooms. The results are shown in the first four columns of table 4. Assigning the best teachers to largest classrooms (counterfactual 1) has a both a positive effect on the mean of test scores and a reduction of inequality. Since teachers are a public good and all students equally benefit from them, assigning better teachers to larger classrooms means that more students can benefit from them, and less students are assigned to low quality teachers. This way, good teachers partially offset the negative effect coming from being in large classrooms.

Table 4: Counterfactuals, mathematics test scores

Counterfactual	(1)	(2)	(3)	(4)	(5)	(6)
mean	0.03	0.04	0.07	0.00	0.03	0.02
sd	-0.01	1.34	1.20	-0.09	-0.02	-0.03
p10	0.06	-1.55	-1.33	0.21	0.06	0.02
p25	0.04	-0.90	-0.73	0.05	0.04	0.02
p50	0.03	-0.09	0.05	-0.10	0.03	0.03
p75	0.03	0.72	0.63	-0.13	0.04	0.03
p90	0.02	1.79	1.61	-0.06	0.02	0.01

Notes: The first row of the table shows the change in the mean test scores with respect to the baseline case, the second row shows the change in the standard deviation, and the last five rows show the change in the test scores for a selected number of percentiles.

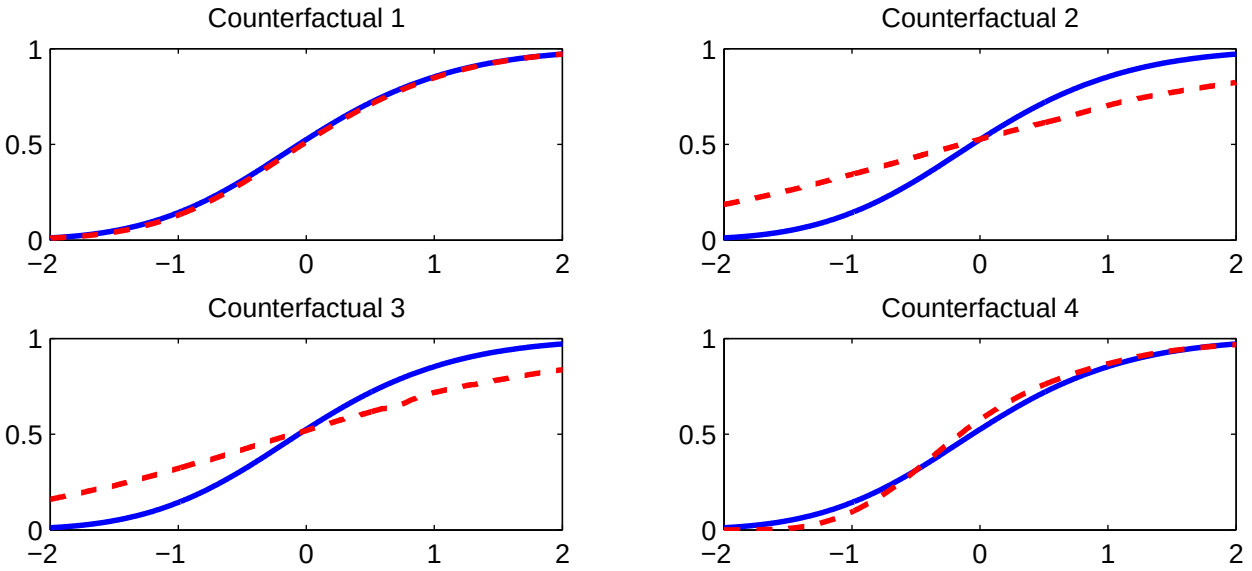
Tracking of students (counterfactual 2) has a positive effect on mean test scores. Since the variance of the student effects is smaller in large classrooms, the student effect is not so negative for bad students, but the good ones, who are assigned to smaller classrooms, have a more positive value, resulting in an overall increase of test scores. This assignment rule reduces the within variance, as students in the same classroom tend to be more similar, but it greatly increases the between classroom variance, resulting in an overall increase of inequality. The combination of the two policies (counterfactual 3) leads to a greater increase of mean test scores, but again the inequality in increased. This result is driven by the peer effects, which dominate the variance-reduction effect of the teacher asignment.[33]

Finally, *detracking* barely affects the mean, but reduces the inequality, as the between variance of the test scores is greatly reduced. This type of matching is particularly effective for students in the lower tail of the distributions, who benefit more from being in the classroom with the best students.

5.2 Changing the Distribution of Class Sizes

Suppose that a principal observes the quality of their teachers, but the ability of their students is unknown. This is a plausible assumption for kindergarten students with whom the principal had no prior interaction. Since students ability is unknown, they are randomly

Figure 3: Distribution of Test Scores



Notes: The solid line represents the distribution of test scores with random assignment into classrooms of both teachers and students, and the solid line represents the same distribution for the four different counterfactuals considered in the text. The distribution of class sizes is the empirical distribution. The estimates used for the counterfactuals are those from specification 6.

assigned into classrooms, but the principal can still affect their test scores by determining how many students each teacher is assigned. If the objective is to maximize the expected average outcome, the maximization problem is the following:

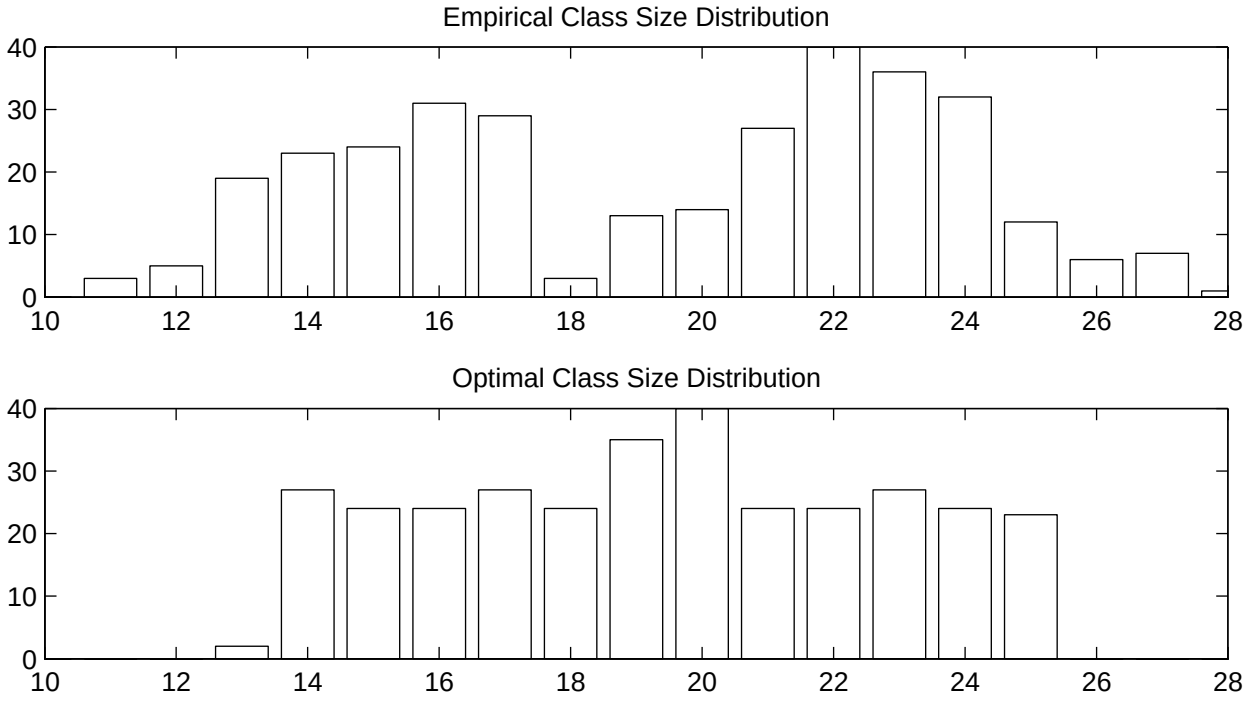
$$(N_1, ..., N_C) = \arg \max_{n_1,...n_C} \frac{1}{N} \sum_{c=1}^{C} \mathbb{E}\left(y_{ic} | n_c, \alpha_c\right) n_c \tag{5}$$

subject to the restriction that all students are assigned to a classroom: $\sum_{j=1}^{N_j} n_j = N$. Conditional on class size and teacher's quality, the expected value of students ability is zero, so $\mathbb{E}\left(y_{ic} | N_c, \alpha_c\right) = \alpha_c\left(N_c\right) = \alpha_{0,c} + \alpha_1 N_c$. This implies that the intercept is different for different teachers, but the slope is the same. The maximum is attained at $N_c = \frac{N}{C} + \frac{1}{2C} \sum_{d=1}^{C} \frac{\alpha_{0,d} - \alpha_{0,c}}{\alpha_1}$ for $c = 1, ..., C-1$ and $N_C = N - \sum_{c=1}^{C-1} N_c$. In words, the better the teacher, the more students he is assigned. This rule trades off the increase in test scores of assigning more students to good teachers with the decrease caused by the increase in class size. Hence, if teachers were equally effective, it would be optimal to have all classes of the same size.

Figure 4 shows the actual class sizes distribution and the optimal distribution using the

estimates of the distribution of teacher effects. The optimal distribution takes values between 13 and 25, in contrast with the observed distribution, which takes values on a broader range. Moreover, the optimal one is quite evenly distributed, being very close to a discrete uniform distribution. Counterfactual 5 in table 4 shows the difference between using the optimal class size rule and the observed class size rule. Since the optimal class size rule assigns more students to the classrooms taught by good teachers, it is not surprising that the results are very similar. Counterfactual 6 shows the effect of reducing the class size dispersion to the minimum, *i.e.* having all class sizes equal to the average number of students per class. This policy would raise the mean performance of the classroom, while at the same time reducing the overall inequality, but relative to the previous counterfactual, the effects would be smaller in magnitude.

Figure 4: Class size distribution



Notes: Absolute frequency of the observed and optimal class size distributions.

5.3 Discussion

One concern with the counterfactuals shown in this paper is that teachers, students, and even parents and principals could react to them, affecting the distribution of outcomes. For example, assignment of good teachers to large classrooms could be interpreted as a reward to bad teachers, thus creating an incentive for teachers to exert less effort.[34] Moreover, these counterfactuals could suffer from misspecification of the model of social interactions. This would be particularly concerning for the counterfactuals that involve tracking of students, since the linearity assumption does not allow for endogenous groups formation within the classroom (Carrell et al., 2013), and nonlinear spillovers could lead to a substantially different counterfactual (Sacerdote, 2011). On the other hand, the counterfactuals regarding the change in the distribution of class sizes, or the assignment of teachers would be more robust to such misspecification.

However, these counterfactuals are useful from a policy perspective: they suggest that, at least for students in kindergarten, in order to decrease overall inequality, measures aiming at reducing the between class inequality tend to be more successful than those aiming at reducing the within class inequality. Also, even though this study is restricted to kindergarten students, it has been found that kindergarten test scores are highly correlated with different economic outcomes during adulthood, such as earnings, college attendance, home ownership, and retirement savings (Chetty et al., 2011).

6 Conclusion

In this paper I propose a method to jointly estimate the strength of peer effects along with several features of the distributions of teacher and student effects. This method uses excess covariance (and higher order moments) analysis and I applied it to the estimation of spillovers in the kindergarten. The results show evidence of strong spillovers in the classroom, with a social multiplier of around 1.5. Moreover, teacher and student effects depart from the

usually maintained normality assumption. In particular, the distribution of teacher effects is thin tailed, *i.e.* there are few extremely effective or ineffective teachers, whereas the student distribution is thick tailed. Increasing the teacher's quality by one standard deviation is associated with an increase in test scores of around 10 to 15% of a standard deviation, and increasing classmates' abilities by one standard deviation is associated with an increase in one's own test scores of around 45% of a standard deviation. Class size has a negative mean effect on test scores, but it also reduces the variance of the student effects, which decreases the overall variance of test scores.

Using the estimation results, I conduct several counterfactual social planning experiments. These experiments show that a resource neutral policy can have a direct impact on the distribution of test scores, with some students benefiting more than others. In particular, assigning good teachers to large classrooms improves the distribution of test scores overall, and reduces inequality simultaneously; tracking of students leads to an increase in test scores for good students, at the cost of a decrease in bad students' test scores; detracking does a good job at reducing the inequality among students without affecting much the mean performance. Finally, I also consider the optimal class size distribution, which assigns more students to better teachers, resulting in a simultaneous increase of the mean test scores and a reduction of the inequality.

References

Altonji, J. G. and L. M. Segal (1996). Small-sample bias in GMM estimation of covariance structures. *Journal of Business & Economic Statistics 14*(3), 353–366.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics 30*, 98–108.

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics 114*(2), 533–575.

Arcidiacono, P., G. Foster, N. Goodpaster, and J. Kinsler (2012). Estimating spillovers using panel data, with an application to the classroom. *Quantitative Economics 3*(3), 421–470.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association 104*(486), 486–500.

Bonhomme, S. and J.-M. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics 149*(1), 12–25.

Bonhomme, S. and J.-M. Robin (2010). Generalized non-parametric deconvolution with an application to earnings dynamics. *The Review of Economic Studies 77*(2), 491–533.

Boucher, V., Y. Bramoullé, H. Djebbari, and B. Fortin (2014). Do peers affect student achievement? Evidence from Canada using group size variation. *Journal of Applied Econometrics 29*(1), 91–109.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*(1), 41–55.

Brock, W. A. and S. N. Durlauf (2001). Interactions-based models. *Handbook of Econometrics 5*, 3297–3380.

Burke, M. A. and T. R. Sass (2013). Classroom peer effects and student achievement. *Journal of Labor Economics 31*(1), 51–82.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer effects and social networks in education. *The Review of Economic Studies 76*(4), 1239–1267.

Carman, K. G. and L. Zhang (2012). Classroom peer effects and academic achievement: Evidence from a chinese middle school. *China Economic Review 23*(2), 223–237.

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

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics 126*(4), 1593–1660.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014, September). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2633–79.

Cragg, J. G. (1997). Using higher moments to estimate the simple errors-in-variables model. *Rand Journal of Economics 28*(0), 71–91.

De Giorgi, G. and M. Pellizzari (2014). Understanding social interactions: Evidence from the classroom. *The Economic Journal 124*(579), 917–953.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics 2*, 241–275.

Durlauf, S. N. and Y. M. Ioannides (2010). Social interactions. *Annual Review of Economics 2*(1), 451–478.

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman (1996). Crime and social interactions. *The Quarterly Journal of Economics CXI*(2), 507–548.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*(3), 643–660.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review 61*(2), 280–288.

Hanushek, E. A. and S. G. Rivkin (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review 100*(2), 267–271.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010). The rate of return to the highscope perry preschool program. *Journal of public Economics 94*(1), 114–128.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics 115*(4), 1239–1285.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics 116*(3), 777–803.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association 78*(381), 47–55.

Murphy, R. and F. Weinhardt (2014). Top of the class: The importance of ordinal rank. Technical report, CESifo Group Munich.

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

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*(2), 247–252.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy 4*(4), 537–571.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? *Handbook of the Economics of Education 3*, 249–277.

Tincani, M. M. (2014). Heterogeneous peer effects and rank concerns: Theory and evidence. Working paper.

Todd, P. and K. I. Wolpin (2014). Estimating a coordination game in the classroom. Technical report, Working Paper.

Word, E. R. et al. (1990). The State of Tennessee's Student/Teacher Achievement Ratio (STAR) project: Technical report (1985-1990). Technical report, Nashville: Tennessee State Department of Education.

Appendix

A A Model of Social Interactions in the Classroom

The model is a simultaneous game of complete information in which both the teacher and the students observe the number of students in class, their individual ability, and teacher's quality. Let individual test scores be determined by the following Cobb-Douglas production function

$$y_{ic} = \exp\left(\zeta_{tc} + \xi_{ic}\right) e_{tc}^{\phi} e_{ic}^{\beta} \tag{6}$$

That is, student i in class c's test score is a function that depends positively on teacher quality, ζ_{tc}, their own student ability, ξ_{ic}, teacher effort and their own student effort. Assume that $\phi < 1$ and $\beta < 1$, so that teacher and student effort are complements in the production function but their marginal returns are decreasing.[35]

The first component of the production function, $\exp\left(\zeta_{tc} + \xi_{ic}\right)$, reflects teacher's quality and student's ability, and it is the way heterogeneity is introduced in this model.[36] Further, teacher's quality and student's ability are allowed to depend on class size. Intuitively, both teachers and students can have a different level of productivity for different levels of class size. For example, some teachers can be more effective at teaching small classrooms than large classrooms, as the larger the classroom, the more opportunities for disruptions there are. Similarly, students can perform differently in classrooms of different sizes. In the most general formulation, there are potential outcomes for each different class size, which are drawn from some distribution, $\zeta_{tc} \equiv \zeta_{tc}\left(N_c\right)$ and $\xi_{ic} \equiv \xi_{ic}\left(N_c\right)$, *i.e.* it would be a random coefficients model with multiple dummy variables, one for each class size. Since it is possible that the distribution of potential outcomes varies for different values of class size, affecting the distribution of test scores, it is important to know these distributions for the teacher and student assignment problem. Moreover, this heterogeneity in teacher and student effects implies that the variance and higher order moments of test scores are a function of class size.

Let students' utility function be linear in their test score. Students incur some cost by exerting effort, which is homogeneous for all individuals and increasing in effort.

$$u_i\left(y_{ic}, e_{ic}\right) = y_{ic} - e_{ic}^{\delta} \tag{7}$$

where y_{ic} is the test score of individual i and e_{ic}^{δ} is their cost function. This particular utility function rules out any kind of rank effect, as only the absolute test score matters, and not the performance relative to their classroom peers.[37] I assume that $\delta > \frac{\beta}{1-\phi}$, so that the marginal cost in effort increases faster than its marginal product. Let teachers' utility function be given by

$$u_c\left(\overline{y_c}^{g}, e_{tc}\right) = \overline{y_c}^{g} - e_{tc} \tag{8}$$

That is, teachers utility is linear in the geometric mean of students' grades, and they incur a cost that is also homogeneous for all teachers.[38] Moreover, the marginal cost is constant in effort.

The baseline model equations rule out the direct spillovers among students in the same classroom. The channel for the spillovers in this model is teacher's effort. Given that agents behave rationally, teachers are going to put effort according to the effort choices and ability of all the students in their classroom. Since students optimal effort level is going to depend on teacher's effort, it follows that students' effort and test scores are going to be indirectly influenced by their peers' effort and abilities. Therefore, teachers fulfill two roles in this model: they directly affect students test scores through their quality and effort, and they allow for the existence of peer effects through the effort they optimally exert. It is also relatively simple to generalize the production function such that it incorporates direct peer effects, although it is no longer possible to obtain closed form solutions for the test scores if the game is simultaneous.

A.1 Solution of the Model

This model is solved using standard game theoretic arguments. Start by obtaining students' optimal effort level, given their individual ability, teacher's quality and conditioning on

teacher's effort level

$$e_{ic}^{*}\left(e_{tc}\right)=\arg\max_{e}\exp\left(\zeta_{tc}+\xi_{ic}\right)e_{tc}^{\phi}e^{\beta}-e^{\delta}$$

Taking the derivative with respect to effort, one gets the first order conditions for this problem. Notice that this is a coordination game, since there exist two possible Nash equilibria. In the first one, every student exerts no effort. To solve for the second Nash equilibrium, it is convenient to work with the logarithm of these *foc*, obtaining

$$\log\left(e_{ic}\right)=\frac{1}{\delta-\beta}\log\left(\frac{\beta}{\delta}\right)+\frac{1}{\delta-\beta}\left(\zeta_{tc}+\xi_{ic}\right)+\frac{\phi}{\delta-\beta}\log\left(e_{tc}\right) \tag{9}$$

The best response function indicates that the optimal effort level of a student depends positively on teacher's quality, student's ability and teacher's effort, which follows from the fact that teacher and student effort are complements in the test score production function. The teacher's best response function is obtained by maximizing the following function:

$$e_{tc}^{*}\left(\left\{e_{jc}\right\}_{j=1}^{N_c}\right)=\arg\max_{e}\exp\left(\zeta_{tc}+\overline{\xi}_{c}\right)e^{\phi}\prod_{j=1}^{N_c}e_{jc}^{\frac{\beta}{N_c}}-e$$

Again, taking logs of the *foc* and solving for teacher's log effort yields

$$\log\left(e_{tc}\right)=\frac{1}{1-\phi}\log\left(\phi\right)+\frac{1}{1-\phi}\left(\zeta_{tc}+\overline{\xi}_{c}\right)+\frac{\beta}{1-\phi}\overline{\log\left(e_{tc}\right)} \tag{10}$$

The best response function of teacher's effort shows that they exert more effort the higher their quality, the higher the average ability of their students, and the higher their effort. This best response function is the channel for the spillovers. Since the teacher cares for all their students, he exerts effort according to the ability of all of them. Moreover, the more effort the teacher exerts, the more effort their students exert, which implies that teacher's effort is a public good from which all students benefit. In equilibrium, effort levels equal

$$\log\left(e_{tc}^{*}\right)=\frac{\delta-\beta}{\delta\left(1-\phi\right)-\beta}\log\left(\phi\right)+\frac{\beta}{\delta\left(1-\phi\right)-\beta}\log\left(\frac{\beta}{\delta}\right)+\frac{\delta}{\delta\left(1-\phi\right)-\beta}\left(\zeta_{tc}+\overline{\xi}_{c}\right) \tag{11}$$

$$\begin{aligned}\log\left(e_{ic}^{*}\right)=&\frac{\phi}{\delta\left(1-\phi\right)-\beta}\log\left(\phi\right)+\frac{1-\phi}{\delta\left(1-\phi\right)-\beta\}}\log\left(\frac{\beta}{\delta}\right)\\&+\frac{1}{\delta\left(1-\phi\right)-\beta}\zeta_{tc}+\frac{\phi\delta}{\left(\delta\left(1-\phi\right)-\beta\right)\left(\delta-\beta\right)}\overline{\xi}_{c}+\frac{1}{\delta-\beta}\xi_{ic}\end{aligned} \tag{12}$$

The optimal student effort levels already take into account the indirect spillovers that there are among them, and thus they depend on four different terms: a constant, teacher's quality, the average ability of their peers, and their own individual ability. Teacher's optimal

effort level is similar, and it depends on a constant, his own quality, and the mean of students' ability. Plugging 11 and 12 into 6 yields the individual test score in equilibrium:

$$\log(y_{ic}) = \frac{\phi\delta}{\delta(1-\phi)-\beta}\log(\phi) + \frac{\beta}{\delta(1-\phi)-\beta}\log\left(\frac{\beta}{\delta}\right) + \frac{\delta}{\delta(1-\phi)-\beta}\zeta_{tc} + \frac{\phi\delta^2}{(\delta(1-\phi)-\beta)(\delta-\beta)}\overline{\xi}_c + \frac{\delta}{\delta-\beta}\xi_{ic} \tag{13}$$

In equilibrium, student's i test score depends positively on teacher's quality, ζ_{tc}, the average ability of the students in class c, $\overline{\xi}_c$, and his own individual ability, ξ_{ic}. It is convenient to rewrite equation 13 as

$$\log(y_{ic}) = \alpha_c + (\gamma-1)\overline{\varepsilon}_c + \varepsilon_{ic} \tag{14}$$

where

$$\alpha_c \equiv \frac{\phi\delta}{\delta(1-\phi)-\beta}log(\phi) + \frac{\beta}{\delta(1-\phi)-\beta}log\left(\frac{\beta}{\delta}\right) + \frac{\delta}{\delta(1-\phi)-\beta}\zeta_{tc}$$

$$\varepsilon_{ic} \equiv \frac{\delta}{\delta-\beta}\xi_{ic}$$

$$\gamma \equiv \frac{\delta-\beta}{\delta(1-\phi)-\beta}$$

That is, I redefine the teacher effect as the sum of the constant and teacher's quality, scaled by $\frac{\delta}{\delta(1-\phi)-\beta}$; the student effect is redefined as the student ability, scaled by $\frac{\delta}{\delta-\beta}$; and gamma is interpreted as the social multiplier, *i.e.* by how much the student test scores would increase if I increased the average student effect by one unit. In terms of the model primitives, the social multiplier equals one when $\phi = 0$. This is the case in which teacher's behavior plays no role, and the production function simplifies to $y_{ic} = \exp(\zeta_{tc} + \xi_{ic})\, e_{ic}^{\beta}$. This implies that teacher's strategic choice of effort, which depends on all students' abilities, has no effect on students' outcomes and therefore students do not benefit from having better peers. Notice that even in this case students benefit from teacher's quality, ζ_{tc}. If $\phi < 1$, better peers have a positive spillover through the increase in teacher's optimal effort.

A.2 Multiplicity of Equilibria

As mentioned above, there are two Nash equilibria that solve the previous model: in the first equilibrium all agents exert no effort; in the second, all agents exert the optimal level of effort

given by equations 11 and 12. The focus of this paper is not to consider a coordination game as in Todd and Wolpin (2014). They have a richer model that also includes the equilibrium in which no agent exerts effort, to which they refer as the trivial equilibrium. As in their paper, I rule out this equilibrium. One compelling reason for this is that if the model were correct, then in classes in which this equilibrium occurred, everyone would have a zero in their test score, which is not observed in the data.

B Operator *vech*

Let A_N be a d-dimensional array with all dimensions of size N. The operator *vech* selects some of the elements of this array and arranges them into a vector. If A_N is a matrix, it selects the diagonal and upper diagonal elements and arrange them row by row:

$$vech\left(A_N\right) = \left(a_{11}, a_{12}, ..., a_{1N}, a_{22}, ..., a_{2N}, ..., a_{NN}\right)'$$

More generally, for d-dimensional arrays it selects the elements $(i_1, i_2, ..., i_d)$ such that $i_1 \leq i_2 \leq ... \leq i_d$ and arrange them lexicographically by dimensions. The size of the resulting vector equals the total number of combinations with repetition, $\binom{N+d-1}{d}$.

C Identification with Missing Test Scores

As mentioned in section 4, not all of the test scores were observed. Therefore, I extend Bonhomme and Robin (2009) to accommodate missing test scores at random. Let N_{0c} denote the number of students in a class, and N_{1c} the number of students whose test scores are observed. Then, Y_c is a vector of dimension N_{1c}, and X_c is a vector of dimension N_{0c}, and the relation between the two of them is given by $Y_c = \Lambda\left(\gamma; N_{0c}, N_{1c}\right) X_c$, where $\Lambda\left(\gamma; N_{0c}, N_{1c}\right) = \left(\iota_{N_{1c}}, \left(I_{N_{1c}}, 0_{N_{1c}} 0'_{N_{0c}-N_{1c}}\right) + \frac{\gamma-1}{N_{0c}} \iota_{N_{1c}} \iota'_{N_{0c}}\right)$. Most of the analysis remains the same, but now the ω_Y^r vectors are smaller, and the Λ_r matrix are also different, as shown in appendix F.

D Cumulants and Cumulant Generating Functions

Let X be a random variable. Its cumulant generating function (CGF), $g_X(t)$, is defined as the logarithm of the moment generating function:

$$g_X(t) \equiv \log\left(\mathbb{E}\left[\exp(X)\right]\right)$$

To obtain the cumulant of order R, simply take the Rth derivative of the CGF with respect to t and evaluate at $t = 0$:

$$\kappa_R(X) \equiv \left.\frac{\partial^R g_X(t)}{\partial t^R}\right|_{t=0}$$

There is a bijection between cumulants and moments. For example, cumulants up to order 4 are $\kappa_{X1} = \mathbb{E}[X]$, $\kappa_{X2} = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$, $\kappa_{X3} = \mathbb{E}\left[(X - \mathbb{E}(X))^3\right]$, and $\kappa_{X4} = \mathbb{E}\left[(X - \mathbb{E}(X))^4\right] - 3\mathbb{E}\left[(X - \mathbb{E}(X))^2\right]^2$. They satisfy the following two properties: let a be a scalar, then the Rth order cumulant of aX equals $\kappa_R(aX) = a^R\kappa_R(X)$; let X and Y be two independent random variables, then the Rth cumulant of their sum equals $\kappa_R(X) = \kappa_R(X) + \kappa_R(Y)$. With these two properties, it is possible to obtain closed form expressions for the cumulants of the between and within variables. The CGF of Y_c, g_{Y_c}, is a linear function of the CGF of teacher and student effects, which I define as g_α and g_ε, respectively

$$g_{Y_c}(t|N_c) = g_\alpha\left(\sum_{j=1}^{N_c} t_j|N_c\right) + \sum_{j=1}^{N_c} g_\varepsilon\left(t_j + \frac{\gamma - 1}{N_c}\sum_{h=1}^{N_c} t_h|N_c\right) \tag{15}$$

Let i, j, h and k denote students of class c. To obtain the joint cumulants of the test scores, simply take the *Rth* derivative of the CGF with respect to the different components of the vector t and evaluate it at $t = 0$.

E Characteristic Functions

The characteristic function of the vector of class test scores can be expressed as a function of the characteristic functions of teacher and student effects. Bonhomme and Robin (2010) showed that using the empirical characteristic functions of the observed data, one can recover the characteristic functions of the underlying processes. The framework in this paper is very similar, but there are three differences: several factors are equally distributed, the size of

the ve$Y$ vector varies for different groups, and some of the observations from this vector are missing. The first difference comes from the fact that students are randomly assigned into classes and therefore student effects are treated as coming from the same distribution. Thus, there is extra structure that can be exploited in this paper's framework. The second and the third differences come from the fact that classrooms have a different number of students and some of the test scores are missing.

Assume for the time being that $N_{0c} = N_{1c}$, *i.e.* all students test scores are observed, and drop the 0/1 subscript. The characteristic function of Y_c is given by

$$\varphi_{Y_c}(t|N_c) = \mathbb{E}\left[\exp\left(i\left(\sum_{j=1}^{N_c}\tilde{y}_{jc}t_j\right)\right)|N_c\right] = \varphi_\alpha\left(\sum_{j=1}^{N_c}t_j|N_c\right)\Pi_{j=1}^{N_c}\varphi_\varepsilon\left(t_j + \frac{\gamma-1}{N_c}\sum_{h=1}^{N_c}t_h|N_c\right) \tag{16}$$

The CGF of the vector of observed test scores equals the logarithm of the previous equation, and after taking the second derivative, one gets the following $N_c \times N_c$ matrix:

$$\nabla\nabla^T g_{Y_c}(t|N_c) = g_\alpha''\left(\sum_{j=1}^{N_c}t_j|N_{0c}\right)$$
$$+\sum_{j=1}^{N_c}g_\varepsilon''\left(t_j + \frac{\gamma-1}{N_c}\sum_{h=1}^{N_c}t_h|N_c\right)\left[\left(\frac{\gamma-1}{N_c}\right)^2\iota_{N_c}\iota_{N_c}' + \frac{\gamma-1}{N_c}\left(\Upsilon_{N_c}(j) + \Upsilon_{N_c}(j)'\right) + \Psi_{N_c}(j)\right]$$

where $\Upsilon_{N_c}(j)$ is a $N_c \times N_c$ matrix of zeros except for column j, whose elements equal one, and $\Psi_{N_c}(j)$ is a $N_c \times N_c$ matrix of zeros except for the element (j,j), which equals one. The next step is to apply the *vech* operator to the matrix of second derivatives of the CGF, and express it as the product of a weighting matrix and a vector with the $N_c + 1$ different second derivatives of the CGF of teacher and students effects. Given that students are randomly sorted into classes, I use the extra information coming from the fact that they are independent and identically distributed. To do so, let $t = \tau\iota_{N_c}$, *i.e.* t is no longer any vector, it gives the same weight, $\tau \in \mathbb{R}$, to all test scores. Apply the *vech* operator to the previous expression to obtain

$$vech\left(\nabla\nabla^T g_{Y_c}(\tau\iota_{N_c}|N_c)\right) = Q\begin{bmatrix} g_\alpha''(N_c\tau|N_c) \\ g_\varepsilon''(\gamma\tau|N_c)\end{bmatrix}$$

where $Q \equiv \left(\iota_{\frac{(N_c+1)N_c}{2}}, vech(I_{N_c}) + \frac{(\gamma^2-1)}{N_c}\iota_{\frac{(N_c+1)N_c}{2}}\right)$. Let Q_j^- denote the jth row of matrix

Q^-, the second derivative of the CGF of the teacher and student effects equals

$$g_{\alpha}^{''}(\tau|N_c) = Q_1^- vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{\tau}{N_c}\iota_{N_c}|N_c\right)\right)$$
$$g_{\varepsilon}^{''}(\tau|N_c) = Q_2^- vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{\tau}{\gamma}\iota_{N_c}|N_c\right)\right)$$

And using the fact that α and ε have both mean zero and $g(0) = 0$, it is possible to doubly integrate the previous expressions to obtain the CGF of the teacher and student effects

$$g_{\alpha}(\tau|N_c) = \int_0^{\tau}\int_0^{u} Q_1^- vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{N_c}\iota_{N_c}|N_c\right)\right) dvdu$$
$$g_{\varepsilon}(\tau|N_c) = \int_0^{\tau}\int_0^{u} Q_2^- vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{\gamma}\iota_{N_c}|N_c\right)\right) dvdu$$

All that remains to do is to take the exponential of those two quantities to get the characteristic function of the teacher and student effects

$$\varphi_{\alpha}(\tau|N_c) = \exp\left(\int_0^{\tau}\int_0^{u} Q_1^- vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{N_c}\iota_{N_c}|N_c\right)\right) dvdu\right)$$
$$\varphi_{\varepsilon}(\tau|N_c) = \exp\left(\int_0^{\tau}\int_0^{u} Q_2^- vech\left(\nabla\nabla^T g_{Y_c}\left(\frac{v}{\gamma}\iota_{N_c}|N_c\right)\right) dvdu\right)$$

Notice that in the last two expressions, in order to have the CGF or characteristic function of the teacher and student effects evaluated at τ, I need two different weighting vectors t. In both cases each test score has the same weight, but they are different for the two functions. For the function of the teacher effect the weight has to be equal to $\frac{1}{N_c}$, and for the student effect the weight equals $\frac{1}{\gamma}$. This means that knowledge of γ is required in order to get estimates of the characteristic function of the student effect.

Now consider again the case in which test scores are missing, *i.e.* $N_{0c} \neq N_{1c}$. The vector of second derivatives of the CGF is expressed as

$$vech\left(\nabla\nabla^T g_{Y_c}(\tau\iota_{N_{1c}}|N_{0c})\right) = Q\begin{bmatrix} g_{\alpha}^{''}(N_{1c}\tau|N_{0c}) + \frac{(\gamma^2-1)N_{1c}}{N_{0c}} g_{\varepsilon}^{''}\left(\frac{(\gamma-1)N_{1c}}{N_{0c}}\tau|N_{0c}\right) \\ g_{\varepsilon}^{''}\left(\frac{\gamma N_{1c}+N_{0c}-N_{1c}}{N_{0c}}\tau|N_{0c}\right)\end{bmatrix}$$

Since there are no observations for the test scores of students $N_{1c}+1, ..., N_{0c}$, there is multicollinearity between their effects and the teacher effect, since they affect all the remaining students proportionally to the teacher. This means that an extra step is needed in order to identify the characteristic function of the teacher effect. The CGF of both the

teacher and student effects equal

$$g_{\varepsilon}\left(\tau|N_{0c}\right)=\int_{0}^{\tau}\int_{0}^{u}Q_{2}^{-}vech\left(\nabla\nabla^{T}g_{Y_{c}}\left(\frac{vN_{0c}}{\gamma N_{1c}+\left(N_{0c}-N_{1c}\right)}\iota_{N_{1c}}|N_{0c}\right)\right)dvdu$$

$$\begin{aligned}g_{\alpha}\left(\tau|N_{0c}\right)=\int_{0}^{\tau}\int_{0}^{u}\Bigg[&Q_{1}^{-}vech\left(\nabla\nabla^{T}g_{Y_{c}}\left(\frac{v}{N_{1c}}\iota_{N_{1c}}|N_{0c}\right)\right)\\&-g_{\varepsilon}^{''}\left(\frac{\left(\gamma-1\right)\left(N_{0c}-N_{1c}\right)}{N_{0c}N_{1c}}v\iota_{N_{1c}}|N_{0c}\right)\Bigg]dvdu\end{aligned}$$

That is, the CGF of ε needs a minor correction that involves only the class size and the observed number of test scores, whereas the CGF of α needs a major correction, as the term is now contaminated by the second derivative of the CGF of ε.

F Λ Matrices

F.1 All Test Scores are Observed

$$\Lambda_{2}\left(\gamma;N_{c}\right)\equiv\iota\left(1,\frac{\gamma^{2}-1}{N_{c}}\right)+\left[0,vech\left(\eta_{2,1,2}^{N_{c}}\right)\right]$$

$$\begin{aligned}\Lambda_{3}\left(\gamma;N_{c}\right)\equiv\iota&\left(1,\frac{\left(\gamma-1\right)^{2}\left(\gamma-2\right)}{N_{c}^{2}}\right)+\left[0,\frac{\gamma-1}{N_{c}}vech\left(\eta_{3,1,2}^{N_{c}}+\eta_{3,1,3}^{N_{c}}+\eta_{3,2,3}^{N_{c}}\right)\right]\\&+\left[0,vech\left(\eta_{3,1,2}^{N_{c}}\odot\eta_{3,1,3}^{N_{c}}\right)\right]\end{aligned}$$

$$\begin{aligned}\Lambda_{4}\left(\gamma;N_{c}\right)\equiv\iota&\left(1,\frac{\left(\gamma-1\right)^{3}\left(\gamma-3\right)}{N_{c}^{3}}\right)\\&+\left[0,\frac{\left(\gamma-1\right)^{2}}{N_{c}^{2}}vech\left(\eta_{4,1,2}^{N_{c}}+\eta_{4,1,3}^{N_{c}}+\eta_{4,1,4}^{N_{c}}+\eta_{4,2,3}^{N_{c}}+\eta_{4,2,4}^{N_{c}}+\eta_{4,3,4}^{N_{c}}\right)\right]\\&+\left[0,\frac{\gamma-1}{N_{c}}vech\left(\eta_{4,1,2}^{N_{c}}\odot\eta_{4,1,3}^{N_{c}}+\eta_{4,1,2}^{N_{c}}\odot\eta_{4,1,4}^{N_{c}}+\eta_{4,1,3}^{N_{c}}\odot\eta_{4,1,4}^{N_{c}}+\eta_{4,2,3}^{N_{c}}\odot\eta_{4,2,4}^{N_{c}}\right)\right]\\&+\left[0,vech\left(\eta_{4,1,2}^{N_{c}}\odot\eta_{4,1,3}^{N_{c}}\odot\eta_{4,1,4}^{N_{c}}\right)\right]\end{aligned}$$

where 0 and ι represent vectors of zeros and ones of the appropriate dimension, *i.e.* $\frac{(N_c+1)N_c}{2}$, $\frac{(N_c+2)(N_c+1)N_c}{6}$ and $\frac{(N_c+3)(N_c+2)(N_c+1)N_c}{24}$, respectively. $\eta_{d,e,f}^{N_c}$ is the d-dimensional array whose d dimensions are all of size N_c and all elements zero except for those that are the same in dimensions e and f, $e<f$, which take value one.[39] Finally, $\odot$ is the Hadamard product, *i.e.* the element-wise product of arrays.

F.2 Missing Test Scores at Random

$$\Lambda_2(\gamma; N_{0c}, N_{1c}) \equiv \iota\left(1, \frac{\gamma^2-1}{N_{0c}}\right) + \left[0, vech\left(\eta_{2,1,2}^{N_{1c}}\right)\right]$$

$$\begin{aligned}\Lambda_3(\gamma; N_{0c}, N_{1c}) \equiv\ & \iota\left(1, \frac{(\gamma-1)^2(\gamma-2)}{N_{0c}^2}\right) \\ & + \left[0, \frac{\gamma-1}{N_{0c}} vech\left(\eta_{3,1,2}^{N_{1c}} + \eta_{3,1,3}^{N_{1c}} + \eta_{3,2,3}^{N_{1c}}\right)\right] + \left[0, vech\left(\eta_{3,1,2}^{N_{1c}} \odot \eta_{3,1,3}^{N_{1c}}\right)\right]\end{aligned}$$

$$\begin{aligned}\Lambda_4(\gamma; N_{0c}, N_{1c}) \equiv\ & \iota\left(1, \frac{(\gamma-1)^3(\gamma-3)}{N_{0c}^3}\right) \\ & + \left[0, \frac{(\gamma-1)^2}{N_{0c}^2} vech\left(\eta_{4,1,2}^{N_{1c}} + \eta_{4,1,3}^{N_{1c}} + \eta_{4,1,4}^{N_{1c}} + \eta_{4,2,3}^{N_{1c}} + \eta_{4,2,4}^{N_{1c}} + \eta_{4,3,4}^{N_{1c}}\right)\right] \\ & + \left[0, \frac{\gamma-1}{N_{0c}} vech\left(\eta_{4,1,2}^{N_{1c}} \odot \eta_{4,1,3}^{N_{1c}} + \eta_{4,1,2}^{N_{1c}} \odot \eta_{4,1,4}^{N_{1c}}\right.\right. \\ & \left.\left. + \eta_{4,1,3}^{N_{1c}} \odot \eta_{4,1,4}^{N_{1c}} + \eta_{4,2,3}^{N_{1c}} \odot \eta_{4,2,4}^{N_{1c}}\right)\right] + \left[0, vech\left(\eta_{4,1,2}^{N_{1c}} \odot \eta_{4,1,3}^{N_{1c}} \odot \eta_{4,1,4}^{N_{1c}}\right)\right]\end{aligned}$$

G Testing for the Presence of Peer Effects

Let the test score of student i in classroom c by given by the following equation:

$$y_{ic} = \alpha_c + (\gamma - 1) h(\varepsilon_{ic}, \varepsilon_{-ic}) + \varepsilon_{ic} \tag{17}$$

As in equation 1, α_c and ε_{ic} respectively represent the teacher and student effects. The only difference lies in the peer effects function, $h(\varepsilon_{ic}, \varepsilon_{-ic})$, which depends on the own student effect, and the effect of all other students in the classroom, ε_{-ic}. This function is left unspecified, and coincides with the linear-in-means model when $h(\varepsilon_{ic}, \varepsilon_{-ic}) = \overline{\varepsilon}_c$, but it can also represent other peer effects models, such as the bad apple model ($h(\varepsilon_{ic}, \varepsilon_{-ic}) = \min_j \varepsilon_{jc}$).

Consider the mean test score of classroom c, $\overline{y}_c = \alpha_c + (\gamma - 1)\frac{1}{N_c}\sum_{i=1}^{N_c} h(\varepsilon_{ic}, \varepsilon_{-ic}) + \overline{\varepsilon}_c$, and the difference of individual test scores with respect to it, $y_{ic} - \overline{y}_c = (\gamma - 1)(h(\varepsilon_{ic}, \varepsilon_{-ic})) +$

$\varepsilon_{ic} - \overline{\varepsilon}_c$. The variance of these two quantities are given by

$$\begin{aligned} Var\left(\overline{y}_c\right) = & Var\left(\alpha_c\right) + \left(\gamma-1\right)^2 \left[\frac{1}{N_c} Var\left(h\left(\varepsilon_{ic}, \varepsilon_{-ic}\right)\right) + \frac{N_c - 1}{N_c} Cov\left(h\left(\varepsilon_{ic}, \varepsilon_{-ic}\right), h\left(\varepsilon_{jc}, \varepsilon_{-jc}\right)\right)\right] \\ & + \frac{1}{N_c} Var\left(\varepsilon_{ic}\right) \\ & + 2\left(\gamma-1\right)\left[\frac{1}{N_c} Cov\left(\varepsilon_{ic}, h\left(\varepsilon_{ic}, \varepsilon_{-ic}\right)\right) + \frac{N_c - 1}{N_c} Cov\left(\varepsilon_{ic}, h\left(\varepsilon_{jc}, \varepsilon_{-jc}\right)\right)\right] \end{aligned} \tag{18}$$

$$\begin{aligned} Var\left(y_{ic} - \overline{y}_c\right) = & \left(\gamma-1\right)^2 \left[\frac{N_c - 1}{N_c} Var\left(h\left(\varepsilon_{ic}, \varepsilon_{-ic}\right)\right) - \frac{N_c - 1}{N_c} Cov\left(h\left(\varepsilon_{ic}, \varepsilon_{-ic}\right), h\left(\varepsilon_{jc}, \varepsilon_{-jc}\right)\right)\right] \\ & + \frac{N_c - 1}{N_c} Var\left(\varepsilon_{ic}\right) \\ & + 2\left(\gamma-1\right)\left[\frac{N_c - 1}{N_c} Cov\left(\varepsilon_{ic}, h\left(\varepsilon_{ic}, \varepsilon_{-ic}\right)\right) - \frac{N_c - 1}{N_c} Cov\left(\varepsilon_{ic}, h\left(\varepsilon_{jc}, \varepsilon_{-jc}\right)\right)\right] \end{aligned} \tag{19}$$

where all the moments are conditional on class size, but for notational simplicity it is omitted in the equations. After some algebra, one obtains

$$\begin{aligned} Var\left(\overline{y}_c\right) - \frac{1}{N_c - 1} Var\left(y_{ic} - \overline{y}_c\right) - Var\left(\alpha_c\right) = & \left(\gamma-1\right)^2 Cov\left(h\left(\varepsilon_{ic}, \varepsilon_{-ic}\right), h\left(\varepsilon_{jc}, \varepsilon_{-jc}\right)\right) \\ & + 2\left(\gamma-1\right) Cov\left(\varepsilon_{ic}, h\left(\varepsilon_{jc}, \varepsilon_{-jc}\right)\right) \end{aligned}$$

Under the null hypothesis of no spillovers, $H_0 : \gamma = 1$, the right hand side of the previous equation equals zero, whereas under the alternative hypothesis of spillovers, $H_0 : \gamma > 1$, it is in general different from zero. Hence, one can construct a test statistic with the between and within class variances, denoted by T_C:

$$T_C \equiv \frac{1}{C}\sum_{c=1}^{C}\left(\overline{y}_c - \hat{\mu}\left(N_c\right)\right)^2 - \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c\left(N_c-1\right)}\sum_{i=1}^{N_c}\left(y_{ic} - \overline{y}_c\right)^2 - \hat{\sigma}_\alpha^2 \tag{20}$$

Under $H_0 : \gamma = 1$, the limiting distribution of the previous test is $\mathcal{N}\left(0, r'V^*r\right)$, where $r \equiv \left(-1, 1, -1\right)'$, and

$$\sqrt{C}\left(\begin{bmatrix} \hat{\sigma}_\alpha^2 \\ \frac{1}{C}\sum_{c=1}^{C}\left(\overline{y}_c - \hat{\mu}\left(N_c\right)\right)^2 \\ \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c\left(N_c-1\right)}\sum_{i=1}^{N_c}\left(y_{ic} - \overline{y}_c\right)^2 \end{bmatrix} - \begin{bmatrix} Var\left(\alpha_c\right) \\ \mathbb{E}\left[Var\left(\overline{y}_c\right)\right] \\ \mathbb{E}\left[\frac{1}{N_c-1} Var\left(y_{ic} - \overline{y}_c\right)\right] \end{bmatrix}\right) \xrightarrow{d} \mathcal{N}\left(0, V^*\right)$$

Some comments are in order: first, both the between and the within variances need to be corrected by the missing test scores, as in Graham (2008); second, one could also use higher order moments of the mean class test scores and their within class variance, but the resulting equations would be algebraically cumbersome even without missing data, which

would further complicate them; third, estimation of V^* is not straightforward, and therefore I use the bootstrap to estimate it.[40]

The results are shown in table 5, and they show that for the mathematics exams, the null hypothesis of no peer effects cannot be accepted at the 95% confidence level for all nine specifications.

Table 5: Tests for the presence of spillovers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
T_C	0.095	0.058	0.056	0.097	0.060	0.058	0.096	0.060	0.058
$c_{0.95}$	0.074	0.053	0.055	0.074	0.052	0.055	0.074	0.052	0.055
T_C	0.094	0.066	0.056	0.092	0.065	0.057	0.087	0.060	0.052
$c_{0.95}$	0.086	0.069	0.069	0.082	0.065	0.064	0.081	0.064	0.062

Notes: Columns 1-9 represent use the estimate of the variance of α_c from the 9 specifications presented in the main text. T_C and $c_{0.95}$ respectively denote the value of the test and the 95% critical value.

H Extra Results

H.1 Heterogeneous Teacher Effects

H.2 Estimation when Teacher's aide affects the Variance

Table 2 showed that classes in which there was a full time teacher's aide had a slightly smaller performance, only barely significant for the reading test, but this variable was excluded from the main covariance analysis. Table 7 shows the estimates of the social multiplier, the teacher and the students' variances when the teacher's aide is allowed to have an impact on the variance of the teacher effect. Having a teacher's aide only marginally increases this variance, though this effect is not significant in any of the three specifications. Moreover, the estimates of the social multiplier are barely changed by the inclusion of this variable in the regression.

Table 6: Heterogeneous teacher effects

	Mathematics	Reading
$\hat{\gamma}$	0.739	0.715*
	(1.142)	(0.392)
$\hat{\kappa}_2\,(\alpha_c \vert small)$	0.123	0.114***
	(0.076)	(0.030)
$\hat{\kappa}_2\,(\alpha_c \vert large)$	0.079	0.066***
	(0.060)	(0.022)
$\hat{\kappa}_3\,(\alpha_c \vert small)$	0.059*	0.054
	(0.032)	(0.035)
$\hat{\kappa}_3\,(\alpha_c \vert large)$	$3.1 \cdot 10^{-4}$	0.004
	(0.011)	(0.011)
$\hat{\kappa}_4\,(\alpha_c \vert small)$	-0.021	0.002
	(0.040)	(0.055)
$\hat{\kappa}_4\,(\alpha_c \vert large)$	-0.084***	-0.0723***
	(0.009)	(0.040)
$\hat{\kappa}_2\,(\varepsilon_{ic} \vert small)$	0.676***	0.703***
	(0.024)	(0.040)
$\hat{\kappa}_2\,(\varepsilon_{ic} \vert large)$	0.783***	0.779***
	(0.040)	(0.059)
$\hat{\kappa}_3\,(\varepsilon_{ic} \vert small)$	0.319***	1.187***
	(0.119)	(0.265)
$\hat{\kappa}_3\,(\varepsilon_{ic} \vert large)$	0.213***	0.934***
	(0.058)	(0.148)
$\hat{\kappa}_4\,(\varepsilon_{ic} \vert small)$	1.104**	4.608***
	(0.488)	(1.524)
$\hat{\kappa}_4\,(\varepsilon_{ic} \vert large)$	0.216	3.196***
	(0.142)	(0.778)

Notes: Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels.

H.3 Estimation Constraining the Variances to be Positive

The results shown in section 4.3 for the case in which the variance of student's ability is homoskedastic were not coherent, as the estimate of the variance of teacher's quality was negative. In this section I present the estimates of the social multiplier when the teacher variance is constrained to be non-negative. Figure 5 shows the estimates of the social multiplier and the value of the objective function for different values of σ_α^2 for the mathematics exam. If the variance is constrained to be weakly positive, then the constrained estimates yield a social multiplier of 1.76. If, on the other hand, the constraint is that the variance be strictly positive, then there are no proper estimates, since the objective function is increasing at zero, and hence for any positive value of the variance, there is always a smaller value such that the objective function evaluated at that value is smaller. In any case, the estimate of the social multiplier is smaller and closer to the estimates when the variance of student's ability is allowed to be heteroskedastic, but still higher. As the variance

Table 7: Teacher's aide affects the variance

	(1)	(4)	(7)
$\hat{\gamma}$	1.874***	1.578***	1.538***
	(0.360)	(0.288)	(0.300)
$\hat{\kappa}_2\,(\alpha_c)$	-0.016	0.018	0.024
	(0.048)	(0.034)	(0.034)
AIDE	0.003	0.007	0.003
	(0.021)	(0.021)	(0.021)
$\hat{\kappa}_2\,(\varepsilon_{ic})$	0.709***	-	-
	(0.021)		
$\hat{\kappa}_2\,(\varepsilon_{ic}\|small)$	-	0.792***	-
		(0.042)	
$\hat{\kappa}_2\,(\varepsilon_{ic}\|large)$	-	0.672***	-
		(0.024)	
$\hat{\mu}_{\varepsilon,2,0}$	-	-	0.942**
			(0.468)
$\hat{\mu}_{\varepsilon,2,1}$	-	-	-0.011
			(0.049)
$\hat{\mu}_{\varepsilon,2,2}$	-	-	0.000
			(0.001)

Notes: Standard errors in parentheses. *, ** and *** denote significant at the 90, 95 and 99 percent levels.

of teacher's quality increases, the social multiplier decreases, obtaining an estimate of the social multiplier of 1.55 when the teacher's quality variance equals 0.025. The results are robust to the inclusion of higher order moments in the estimation (not shown here).

H.4 Full Results

In this section I present the full table with the estimates of specifications 1 to 9, as described in section 3, for both the mathematics and reading test scores. Moreover, I also present a table with the results of the specification that allows for heterogeneous teacher and student effects. These effects take two different distributions for small and large classrooms.

Figure 5: Social multiplier and objective function as a function of σ_α^2, mathematics test scores

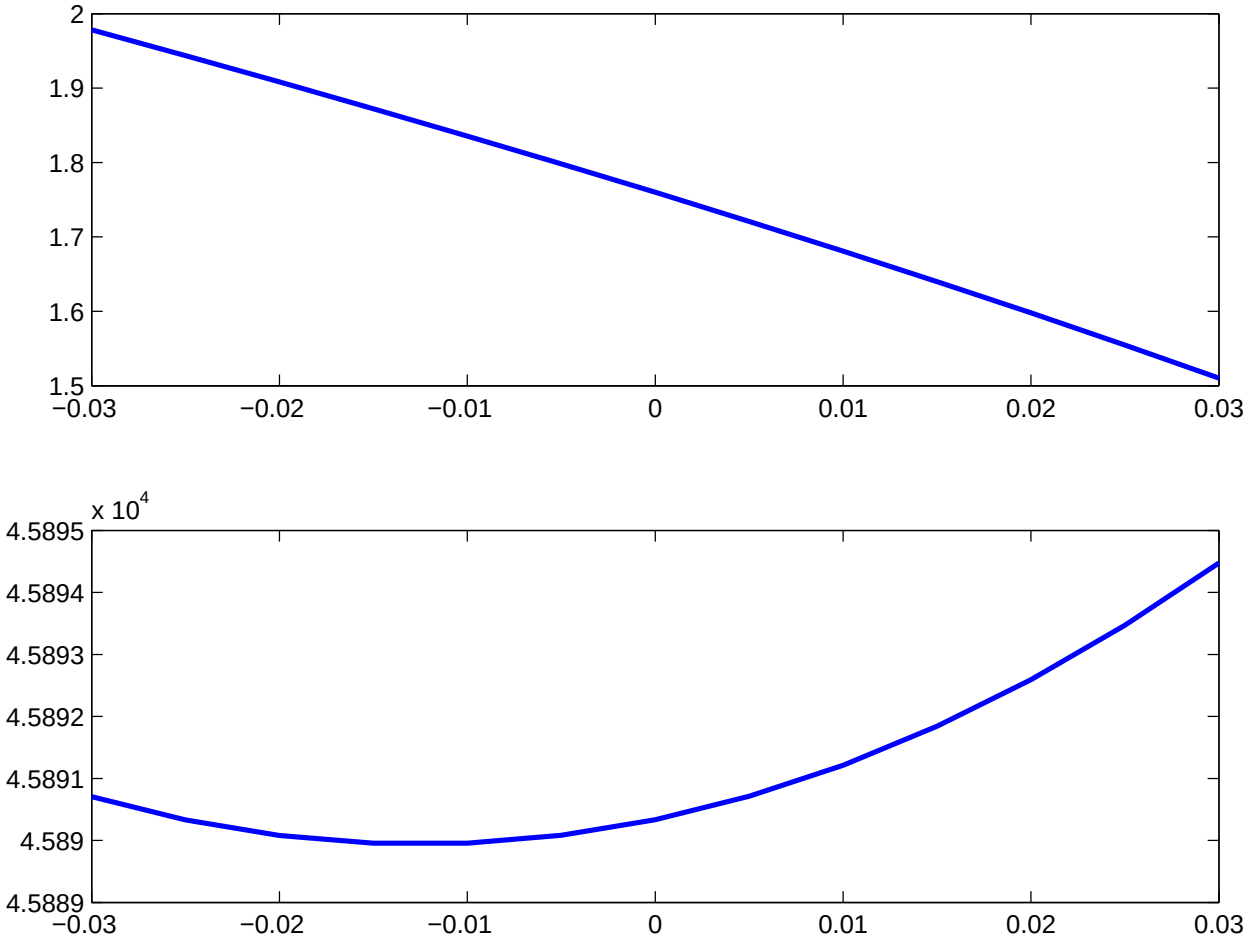


Table 8: Full estimates, mathematics test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$	1.854***	1.868***	1.867***	1.545***	1.564***	1.564***	1.520***	1.544***	1.544***
	(0.374)	(0.395)	(0.374)	(0.299)	(0.299)	(0.300)	(0.311)	(0.311)	(0.312)
$\hat{\kappa}_2(\alpha_c)$	-0.013	-0.014	-0.014	0.024	0.022	0.022	0.027	0.024	0.024
	(0.050)	(0.053)	(0.050)	(0.034)	(0.034)	(0.034)	(0.035	(0.035))	(0.035)
$\hat{\kappa}_3(\alpha_c)$	-	0.007	0.007	-	0.008	0.008	-	0.008	0.008
		(0.012)	(0.010)		(0.010)	(0.010)		(0.010)	(0.010)
$\hat{\kappa}_4(\alpha_c)$	-	-	-0.076***	-	-	-0.075***	-	-	-0.076***
			(0.009)			(0.010)			(0.010)
$\hat{\kappa}_2(\varepsilon_{ic})$	0.709***	0.709***	0.709***	-	-	-	-	-	-
	(0.021)	(0.021)	(0.021)						
$\hat{\kappa}_3(\varepsilon_{ic})$	-	0.242***	0.242***	-	-	-	-	-	-
		(0.083)	(0.049)						
$\hat{\kappa}_4(\varepsilon_{ic})$	-	-	0.263*	-	-	-	-	-	-
			(0.136)						
$\hat{\kappa}_2(\varepsilon_{ic}\|small)$	-	-	-	0.791***	0.789***	0.789***	-	-	-
				(0.042)	(0.041)	(0.041)			
$\hat{\kappa}_2(\varepsilon_{ic}\|large)$	-	-	-	0.672***	0.673***	0.673***	-	-	-
				(0.024)	(0.024)	(0.024)			
$\hat{\kappa}_3(\varepsilon_{ic}\|small)$	-	-	-	-	0.367***	0.367***	-	-	-
					(0.115)	(0.115)			
$\hat{\kappa}_3(\varepsilon_{ic}\|large)$	-	-	-	-	0.187***	0.187***	-	-	-
					(0.057)	(0.057)			
$\hat{\kappa}_4(\varepsilon_{ic}\|small)$	-	-	-	-	-	0.899**	-	-	-
						(0.355)			
$\hat{\kappa}_4(\varepsilon_{ic}\|large)$	-	-	-	-	-	0.014	-	-	-
						(0.124)			

$\hat{\mu}_{\varepsilon,2,0}$	-	-	-	-	-	-	0.933*	0.928***	0.928**
							(0.467)	(0.465)	(0.465)
$\hat{\mu}_{\varepsilon,2,1}$	-	-	-	-	-	-	-0.010	-0.010	-0.010
							(0.049)	(0.049)	(0.049)
$\hat{\mu}_{\varepsilon,2,2}$	-	-	-	-	-	-	$4.5{\cdot}10^{-5}$	$4.5{\cdot}10^{-5}$	$4.7{\cdot}10^{-5}$
							$(1.2{\cdot}10^{-3})$	$(1.2{\cdot}10^{-3})$	$(1.2{\cdot}10^{-3})$
$\hat{\mu}_{\varepsilon,3,0}$	-	-	-	-	-	-	-	-2.281	-2.283
								(5.264)	(5.263)
$\hat{\mu}_{\varepsilon,3,1}$	-	-	-	-	-	-	-	0.438	0.439
								(0.845)	(0.845)
$\hat{\mu}_{\varepsilon,3,2}$	-	-	-	-	-	-	-	-0.023	-0.023
								(0.044)	(0.044)
$\hat{\mu}_{\varepsilon,3,3}$	-	-	-	-	-	-	-	$3.8{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
								$(7.3{\cdot}10^{-4})$	$(7.4{\cdot}10^{-4})$
$\hat{\mu}_{\varepsilon,4,0}$	-	-	-	-	-	-	-	-	-138.55**
									(61.65)
$\hat{\mu}_{\varepsilon,4,1}$	-	-	-	-	-	-	-	-	30.49**
									(13.35)
$\hat{\mu}_{\varepsilon,4,2}$	-	-	-	-	-	-	-	-	-2.429**
									(1.051)
$\hat{\mu}_{\varepsilon,4,3}$	-	-	-	-	-	-	-	-	0.084**
									(0.036)
$\hat{\mu}_{\varepsilon,4,4}$	-	-	-	-	-	-	-	-	-0.001**
									$(4.5{\cdot}10^{-4})$

Table 9: Full estimates, reading test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\hat{\gamma}$	1.791***	1.776***	1.733***	1.553***	1.545***	1.505***	1.466***	1.471***	1.427***
	(0.413)	(0.456)	(0.416)	(0.349)	(0.341)	(0.344)	(0.371)	(0.361)	(0.364)
$\hat{\kappa}_2(\alpha_c)$	-0.020	-0.019	-0.013	0.018	0.009	0.013	0.022	0.017	0.022
	(0.054)	(0.058)	(0.052)	(0.040)	(0.039)	(0.038)	(0.040)	(0.039)	(0.038)
$\hat{\kappa}_3(\alpha_c)$	-	0.001	0.002	-	0.004	0.004	-	0.004	0.005
		(0.014)	(0.011)		(0.011)	(0.011)		(0.010)	(0.011)
$\hat{\kappa}_4(\alpha_c)$	-	-	-0.072***	-	-	-0.070***	-	-	-0.069***
			(0.012)			(0.012)			(0.012)
$\hat{\kappa}_2(\varepsilon_{ic})$	0.727***	0.728***	0.728***	-	-	-	-	-	-
	(0.033)	(0.033)	(0.033)						
$\hat{\kappa}_3(\varepsilon_{ic})$	-	0.882***	0.887***	-	-	-	-	-	-
		(0.146)	(0.125)						
$\hat{\kappa}_4(\varepsilon_{ic})$	-	-	2.730***	-	-	-	-	-	-
			(0.609)						
$\hat{\kappa}_2(\varepsilon_{ic}\|small)$	-	-	-	0.793***	0.793***	0.796***	-	-	-
				(0.060)	(0.059)	(0.059)			
$\hat{\kappa}_2(\varepsilon_{ic}\|large)$	-	-	-	0.697***	0.697***	0.697***	-	-	-
				(0.040)	(0.040)	(0.040)			
$\hat{\kappa}_3(\varepsilon_{ic}\|small)$	-	-	-	-	1.067***	1.075***	-	-	-
					(0.261)	(0.261)			
$\hat{\kappa}_3(\varepsilon_{ic}\|large)$	-	-	-	-	0.835***	0.840***	-	-	-
					(0.141)	(0.142)			
$\hat{\kappa}_4(\varepsilon_{ic}\|small)$	-	-	-	-	-	3.697***	-	-	-
						(1.329)			
$\hat{\kappa}_4(\varepsilon_{ic}\|large)$	-	-	-	-	-	2.567***	-	-	-
						(0.681)			

$\hat{\mu}_{\varepsilon,2,0}$	-	-	-	-	-	-	0.286	0.285	0.288
							(0.743)	(0.743)	(0.749)
$\hat{\mu}_{\varepsilon,2,1}$	-	-	-	-	-	-	0.062	0.062	0.062
							(0.079)	(0.079)	(0.080)
$\hat{\mu}_{\varepsilon,2,2}$	-	-	-	-	-	-	-0.002	-0.002	-0.002
							(0.002)	(0.002)	(0.002)
$\hat{\mu}_{\varepsilon,3,0}$	-	-	-	-	-	-	-	-11.63	-11.63
								(13.90)	(13.18)
$\hat{\mu}_{\varepsilon,3,1}$	-	-	-	-	-	-	-	1.992	1.995
								(2.090)	(2.103)
$\hat{\mu}_{\varepsilon,3,2}$	-	-	-	-	-	-	-	-0.100	-0.100
								(0.108)	(0.109)
$\hat{\mu}_{\varepsilon,3,3}$	-	-	-	-	-	-	-	0.002	0.002
								(0.002)	(0.002)
$\hat{\mu}_{\varepsilon,4,0}$	-	-	-	-	-	-	-	-	-98.76
									(287.19)
$\hat{\mu}_{\varepsilon,4,1}$	-	-	-	-	-	-	-	-	19.38
									(62.26)
$\hat{\mu}_{\varepsilon,4,2}$	-	-	-	-	-	-	-	-	-1.304
									(4.939)
$\hat{\mu}_{\varepsilon,4,3}$	-	-	-	-	-	-	-	-	0.037
									(0.170)
$\hat{\mu}_{\varepsilon,4,4}$	-	-	-	-	-	-	-	-	$3.7{\cdot}10^{-4}$
									(0.002)

H.5 Goodness of Fit

I compare the goodness of fit of the different specifications by looking at the value attained by the objective function at the minimum. This comparison requires the objective function to be the same, so it is possible to compare specifications 3, 6, and 9, because they use cumulants two to four in the estimation, but it is not possible to compare models 7, 8, and 9 because the objective function is the same. Table 10 shows the results. For the mathematics test scores, the model with class type heteroskedasticity for student effects achieves the smallest value of the objective function of all three models, and the random coefficients model in class size for student effect has a similar fit. On the other hand, the model that assumes homoskedastic teacher and student effects does a poorer job than the other two. For the reading test scores the results are similar, but in this case it is the specification with student effects following a random coefficients model attaining the minimum value.

Table 10: Goodness of Fit

	Mathematics test scores			Reading test scores		
	(1)	(2)	(3)	(4)	(5)	(6)
Homoskedasticity	45889.9	55945.3	59273.8	63995.3	88224.6	103950.8
Class type heteroskedasticity	45871.7	55926.5	59254.3	63983.4	88211.1	103934.9
Random coefficients model	45878.7	55933.7	59261.2	63978.7	88203.9	103925.9

Notes: Columns 1 and 4 refer to the estimation with only the variances, columns 2 and 5 refer to the estimation with the variances and third order cumulants, and columns 3 and 6 refer to the estimation with the cumulants up to order four.

The estimates of the third and fourth cumulants are in many cases significantly different from zero. If teacher and student effects were normal, these cumulants would be equal to zero. In that case, the estimates of the variance of the teacher and student effects are sufficient to characterize these distributions. Compare the increase in the fit of the model by looking at the difference in the objective function when using the estimates that assume normality with those that relax this assumption and allow for nonzero third and fourth order cumulants. Table 11 shows the results. Columns 1 and 2 report the value of the objective

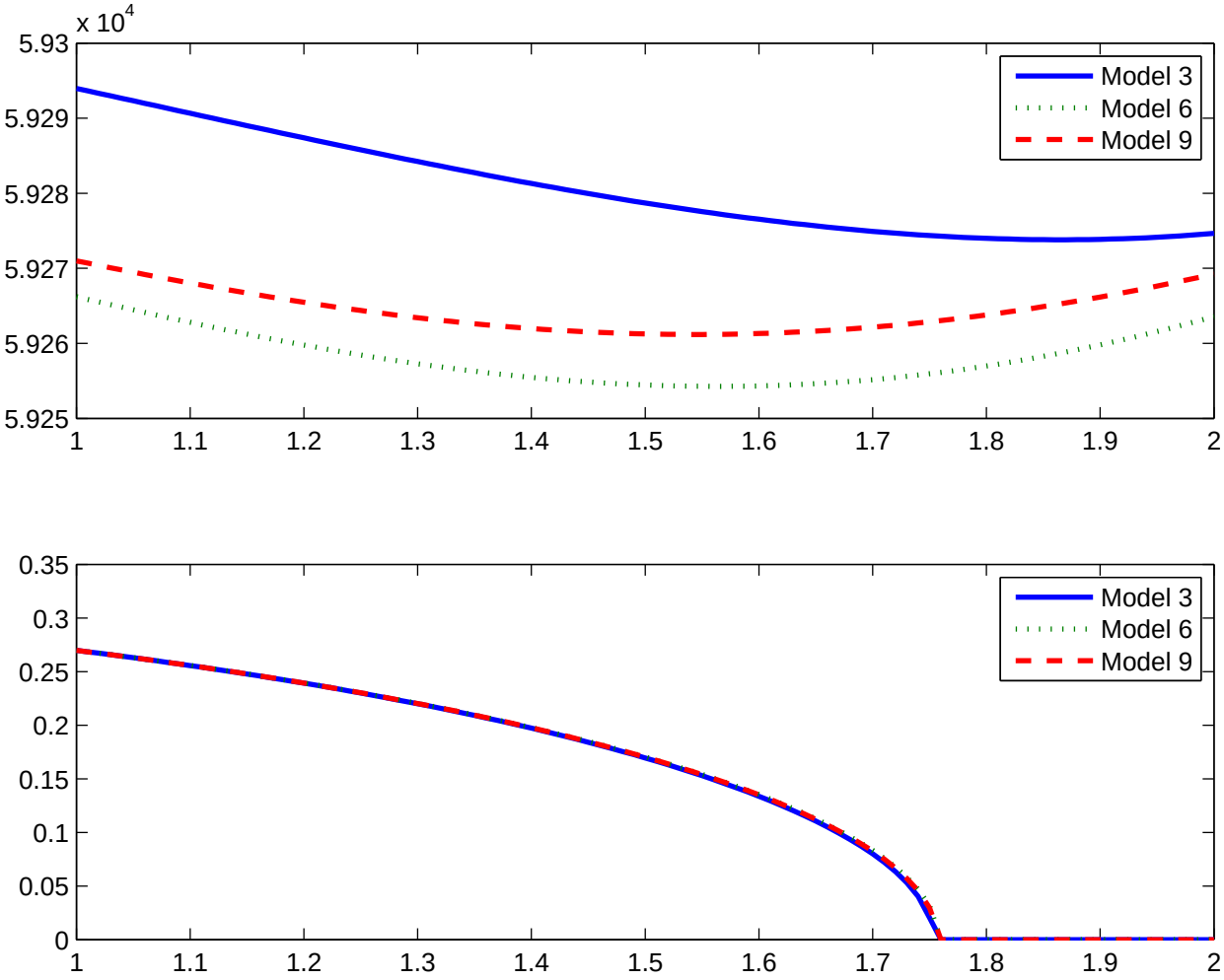
function when using only the second and third cumulants, whereas columns 3 and 4 report the value of the objective function when using the second, third and fourth cumulants.[41] The fit under normality is always worse, and this is particularly marked for the reading test scores.

Table 11: Goodness of fit under normality

	Mathematics test scores			
	Cumulants 2 & 3		Cumulants 2 to 4	
	Non-normality	Normality	Non-normality	Normality
Homoskedasticity	55945.3	55954.8	59273.8	59290.3
Class type heteroskedasticity	55926.5	55936.5	59254.3	59272.0
Random coefficients model	55933.7	55943.6	59261.2	59279.1
	Reading test scores			
	Cumulants 2 & 3		Cumulants 2 to 4	
	Non-normality	Normality	Non-normality	Normality
Homoskedasticity	88224.6	88332.6	103950.8	104093.2
Class type heteroskedasticity	88211.1	88320.8	103934.9	104081.3
Random coefficients model	88203.9	88316.1	103925.9	104076.6

Given that the estimates of the social multiplier are substantially large, one might argue that if the fit of the model remains largely unchanged without spillovers, it would be possible to argue that the estimates of the social multiplier reflect sample variability rather than actual peer effects. The top graph in figure 6 shows the value of the objective function for different values of the social multiplier, and the remaining parameters are the estimates conditional on the social multiplier. The results show that for values of the social multiplier between 1 and 2, the rank in the performance of each model is the same. Hence, the model with homoskedastic teacher and student effects has the poorest fit and the models with heteroskedastic student effects have a better fit. These last two models have a very similar difference in the objective function for each value of the social multiplier, whereas the difference between any of this two and the model with homoskedastic teacher and student effects is decreasing as the social multiplier increases. This is because the estimate of the social multiplier is much larger in the latter model than in the former two.

Figure 6: Goodness of fit and standard deviation of teacher effects as a function of γ, mathematics test scores



Because the sum of the total variance in the test scores is the sum of the variances of student and teacher effects, weighted by the social multiplier, there is a tension between these two estimates: if social multiplier is large, the variance of the teacher effect is small, and the other way around. A particularly interesting case is to restrict the social multiplier to be one, and see what the estimates of the standard deviation of teacher effects are in that case. The bottom graph of figure 6 shows the estimates of the standard deviation of the teacher effect for different values of the social multiplier. For the three models, the estimates of the standard deviation of teacher effects are very close, and if the actual value of the social multiplier were 1, the estimate of the standard deviation of teacher effects would

be approximately 0.27, a number much higher than what has been usually found in the literature. Moreover, for values of the social multiplier larger than 1.75, the estimate of the variance is negative, which suggests that the social multiplier cannot be that large.

H.6 Counterfactuals using the Normal Distribution

Table 12: Counterfactual results, mathematics test scores

Counterfactual	(1)	(2)	(3)	(4)
mean	0.03	0.06	0.09	0.00
sd	-0.01	1.30	1.15	-0.11
p10	0.05	-1.55	-1.33	0.15
p25	0.04	-0.90	-0.75	0.07
p50	0.04	-0.15	-0.11	0.02
p75	0.03	0.61	0.55	-0.06
p90	0.02	1.25	1.15	-0.10

Notes: The first row of the table shows the change in the mean test scores with respect to the baseline case, the second row shows the change in the standard deviation, and the last five rows show the change in the test scores for a selected number of percentiles.

Notes

[1]A major question in the economics of education literature is the estimation of the mean effect of class size on students achievement, but its distributional effects have received less attention. One exception is Lazear (2001), who proposed a model in which the disruptions in a classroom depend on its size, and consequently class size affects the whole distribution of test scores.

[2]See Brock and Durlauf (2001) or Durlauf and Ioannides (2010) for literature reviews of the estimation of spillovers in general, and Sacerdote (2011) for a review of the estimation of peer effects in education.

[3]Morris (1983) method to correct for the bias caused by the incidental parameter problem, which invokes the Gaussianity of the unobserved effects, is typically used for the estimation of teacher value-added (Kane and Staiger, 2008; Chetty et al., 2014). It has already been considered that sorting can bias the teacher value-added estimates (Rothstein, 2009), but not the effect of non-normally distributed teacher effects on these estimates. Rockoff (2004) also assumes normality of teacher effects to estimate its actual distribution.

[4]In this model, students and teachers play a game in which test scores depend on the amount of effort they exert and their ability. The social multiplier arises from the complementarity between students and teacher efforts. Hence, in this framework the outcome is not a choice variable, which is a conceptually different framework from those in which the outcome variable is the choice variable, such as the decision of smoking depending on whether your friends smoke or not. The model links the social multiplier and the moments of the distributions of teacher and student effects to its fundamental parameters, and is solved and discussed in detail in appendix A.

[5]The standard identification strategy to identify peer effects is based on heterogeneous reference groups, which exploit the partial overlap of the group of peers that affect each individual. Bramoullé et al. (2009) formalize this empirical strategy and describe the asymptotic properties of the estimator. Calvó-Armengol et al. (2009), De Giorgi et al. (2010), De Giorgi and Pellizzari (2014), Arcidiacono et al. (2012), and Boucher et al. (2014) use this empirical strategy to estimate the spillovers.

[6]To my knowledge, the earliest example of using covariances to estimate spillovers is Glaeser et al. (1996), though both their framework and the setup are different.

[7]The estimator proposed by Graham (2008) requires a correction of the variances that is complicated to apply to higher order moments. Appendix C shows how this issue is addressed in this paper.

[8]Using all the strength from assumption 1, the characteristic function of Y_c can be expressed as the product of $N_c + 1$ different characteristic functions. See appendix E for the complete derivation.

[9]There is a bijection between cumulants and moments. See appendix D for more details.

[10]In particular, the third cumulant has five different permutations ($i = j = h$, $i = j \neq h$, $i = h \neq j$, $j = h \neq i$, and i, j, h all different) and the fourth cumulant has eighteen different permutations.

[11]The *vech* operator is defined in appendix B; Γ_{Y,N_c} and Ω_{Y,N_c} are the three and four-dimensional arrays that contain all the third and fourth order cumulants of vector Y_c, respectively.

[12]For expositional purposes, consider a student. This model assumes that his effect varies with class size *monotonously*, either increasing if $\varepsilon_{1ic} > 0$ or decreasing if $\varepsilon_{1ic} < 0$. However, different students get different draws of $(\varepsilon_{0ic}, \varepsilon_{1ic})$, which means that some are more efficient at learning in large classes than in small classes and the other way around. This model provides a parsimonious way to capture heterogeneity in teacher and student effects at the class size level.

[13]There are $2H$ for the variances (variance and covariance), $3H$ for the third cumulants (all test scores are of the same student, two are of the same student and the other one is different, or the three of them are of different students), and $5H$ for the fourth cumulants (all test scores are of the same student, three are of the same student and the other one is different, two of them are of the same student and the other two are of a different student, two of them are of the same student and the other two are of different students, or

the four of them are of different students).

[14]If the effects are homoskedastic, $(\gamma, Var(\alpha_c), Var(\varepsilon_{ic}))$; if the effects are class type heteroskedastic, $(\gamma, Var(\alpha_c|type), Var(\varepsilon_{ic}|type))$ for $type = \{small, large\}$; under the random coefficients model in class size for the effects, $(\gamma, Var(\alpha_{0c}), Var(\alpha_{1c}), Cov(\alpha_{0c}, \alpha_{1c}), Var(\varepsilon_{0c}), Var(\varepsilon_{1c}), Cov(\varepsilon_{0c}, \varepsilon_{1c}))$.

[15]In particular, for the second model it includes $\kappa_R(\alpha_c|small)$ and $\kappa_R(\alpha_c|large)$ for the teacher cumulants and similarly for the student cumulants. For the third model, they depend on $\{\mu_{\alpha,R,r}, \mu_{\varepsilon,R,r}\}_{r=0}^{R}$.

[16]To illustrate this, for a classroom of size 18, the number of second, third and fourth order cumulants would be 171, 1140, and 5985, respectively. In relative terms, their weights would be 2%, 16%, and 82%.

[17]These weights are proportional to the variance of the second, third and fourth power of a standard normal random variable. An alternative would be to use the estimated optimal minimum distance weighting matrix. Despite having appealing asymptotic properties, there are two compelling reasons not to use it in this case: as Altonji and Segal (1996) showed, using such matrix when the sample is small would result in biased estimates, particularly for distributions with thick tails; also, because of the number of permutations, the dimension of the weighting matrix would be excessively large to implement the optimal one.

[18]For a more detailed explanation of the experiment and its results, see Word et al. (1990).

[19]Since teachers were always in the same classroom, it is impossible to distinguish between teacher and classroom specific effects. Hereafter I refer to the combined teacher-classroom effect as the teacher effect.

[20]Full randomization took place in the 28 schools (out of 79) in which there was only one class of each type. In the remaining schools, principals could assign teachers and students within classrooms of the same type, but Nye et al. (2004) and Graham (2008) results indicate that using the full sample or only the subsample for which there is fully randomization led to very similar estimates, but more imprecise, so I restrict to the analysis the full sample in this paper.

[21]The length of the experiment was four years, following a cohort from kindergarten to third grade. After the randomization in kindergarten, transfers of students between schools could have created some degree of sorting of students, thus invalidating assumption 2. Therefore, I restrict the analysis to the first year of the experiment.

[22]Following Graham (2008) I assume that the missing test scores are a random sample of the population of test scores. For the mathematics and reading exams, there are 5856 and 5646 observations, respectively.

[23]Alternative specifications allowing the effect of class on test scores to be nonlinear did not yield significantly different results.

[24]The results when I consider heterogeneous teacher effects are shown in appendix H.1. When I include those, the estimates of the teacher cumulants for small and large classes are not significantly different, and the social multiplier is below 1, suggesting some kind of misspecification. If I let the teacher's aide to have

a differential effect on the variance of teacher effect, the results are largely unchanged. See appendix H.2.

[25]I also obtained the estimates without Cragg (1997) weights, obtaining a slightly larger estimate of the social multiplier, but much more imprecisely estimated. Results available upon request.

[26]See appendix H.4 for the tables with all estimates.

[27]For comparison with Graham (2008) estimates, the estimate of the square of the social multiplier ranges between 2.1 and 3.4, which are similar to the estimates he obtained, which were between 2.3 and 3.5.

[28]It is possible to estimate the remaining parameters conditional on σ_α^2 being positive. The results in this case yield a smaller social multiplier, more in line with the estimates of the heteroskedastic models. The details are shown in appendix H.3.

[29]$f_X\left(x;\mu,\sigma,\lambda,\alpha\right) = \left(\sigma\alpha^{\frac{1}{\alpha}-1}\Gamma\left(\frac{1}{\alpha}\right)\right)^{-1} e^{-\left(\frac{|x-\mu|^\alpha}{\sigma^\alpha\alpha}\right)}\Phi\left(sign\left|\frac{x-\mu}{\sigma}\right|\left|\frac{x-\mu}{\sigma}\right|^{\frac{\alpha}{2}}\lambda\left(\frac{2}{\alpha}\right)^{\frac{1}{2}}\right)$ is the pdf of the SEP distribution, where $\Phi\left(\cdot\right)$ is the standard normal cdf and $\Gamma\left(\cdot\right)$ is the gamma function. If $\lambda = 0$ and $\alpha = 2$, then X is a normal distribution with parameters $\left(\mu, \frac{\sigma^2}{2}\right)$.

[30]Bhattacharya (2009) considers both the maximization of students' test scores and other academic outcomes.

[31]Teachers and students may have different potential outcomes for different class sizes, but only the marginal distribution is identified. In these counterfactuals I assume that the rank is the same for all class sizes, which implies that there are no gains by reassigning students to classrooms of a size in which their rank is higher.

[32]To avoid changes in the distribution being driven by a particular school, I take the mean of the school fixed effects and the regular with aide dummy as the intercept of the test scores equation.

[33]The counterfactuals using the normal distribution yielded a similar qualitative answer, but the mean increase was larger and the standard deviation increase was smaller. See table 12 in appendix H.6

[34]This would be a concern on the dynamic performance of teachers, which would not be incompatible with the findings on Projects STAR report (Word et al., 1990), in which it was stated that the pattern of instruction of teachers did not seem to vary with class size.

[35]Mathematically, $\frac{\partial y_{ic}}{\partial e_{ic}} > 0$, $\frac{\partial y_{ic}}{\partial e_{tc}} > 0$, $\frac{\partial y_{ic}}{\partial e_{jc}} = 0$, $\frac{\partial^2 y_{ic}}{\partial e_{ic}^2} < 0$, $\frac{\partial^2 y_{ic}}{\partial e_{tc}^2} < 0$ and $\frac{\partial^2 y_{ic}}{\partial e_{ic}\partial e_{tc}} > 0$.

[36]Teacher's effort and teacher's quality are public goods, as the teacher affects all students equally.

[37]This assumption, while not testable, is arguably likely to be satisfied in this paper's context, because the test was low stakes. Studies finding an effect of the rank on academic achievement (Murphy and Weinhardt, 2014; Tincani, 2014) focused on secondary education students, whose behavior may differ from kindergarten students.

[38]The use of the geometric mean of students' grades is not the most common choice for a utility function. However, since the model is solved in logarithms, and the logarithm of the geometric mean of the test scores

is the arithmetic mean of the logarithm of the test scores, using the geometric mean is convenient.

[39]For example, $\eta^N_{2,1,2} = I_N$, and for the array $\eta^N_{3,1,2}$, its element (i, j, h) equals one if $i = j$, and is zero otherwise. The total number of nonzero elements is N_c^{d-1}.

[40]Because the data is independent at the class level, I draw C classes at random with replacement a total number of $R = 200$ repetitions.

[41]The results when using only the variances in the objective function are the same for both estimators, so they are not reported.