Teacher Instruction, Classroom Composition, and Student Achievement

Andrea Salvati *

(Click here for the most recent version)

November 10, 2023

Abstract

This paper explores teachers' instructional decisions and their implications for the distribution of student achievement. Canonical models of student performance often assume that teacher effectiveness is independent of the classroom environment. In practice, however, teachers can endogenously adapt instruction based on the composition of the classroom. This can have implications for the design of education policies whose impact is likely mediated by teachers' behavior. I exploit unique data from US elementary schools with rich information on teacher instruction to develop and estimate an equilibrium model of endogenous teacher instructional choices, student effort, and student achievement. Teachers are heterogeneous in their teaching ability and choose instructional effort and the allocation of class time across topics. Students vary by initial ability and choose study effort. Student achievement depends on both teacher and student inputs. The model specification allows me to assess whether teachers value unequally the achievement of students with different levels of ability. I find that teachers place a higher value on the achievement of students at the bottom of the ability distribution. I then perform a counterfactual analysis where I reallocate students to classrooms based on prior test score performance (ability tracking) and teachers to classrooms based on teaching ability (assortative matching). Results show that tracking has heterogeneous effects on students with different levels of ability, and that the distribution of these impacts depends on how teachers endogenously adjust their instructional choices to the composition of the classroom. Moreover, the combination of tracking with assigning high-ability teachers to low-ability students would benefit students both at the top and at the bottom of the ability distribution. High-ability students would benefit from spillovers from high-ability peers, while low-ability students would gain from the higher quality and better tailored instruction provided by high-ability teachers.

*Department of Economics, University College London. Email: andrea.salvati@ucl.ac.uk. *Acknowledgements:* I am deeply indebted to Flávio Cunha and Kenneth Wolpin for their guidance and encouragement. I owe special thanks to Rossella Calvi, Gabriella Conti, Isabelle Perrigne, and Michela Tincani for their many comments and support. I also thank Peter Arcidiacono, Pedro Carneiro, Fabrizio Colella, Xiaodong Fan, Jeremy Fox, YingHua He, Qinyou Hu, Ajinkia Keskar, Bolun Li, Lance Lochner, Joseph Mullins, Matthew Thirkettle, Ruth N. López Turley, Daniel Prudencio, Christopher Udry, Andrei Zeleneev and all the seminar and workshop participants at SSES, ESPE, EEA-ESEM, EALE, and SEA conferences, Labor Econometrics Workshop at Monash University, GEEZ seminars; Rice University, University College London, Institute of Fiscal Studies, New York University, USI Lugano, University of Bergen, University of Stavanger, Tec de Monterrey, and Nazarbayev University for useful comments and feedback. All the errors, omissions, and interpretations are my own.

1. Introduction

Disadvantaged students are consistently underachieving in the United States. Recent estimates show that the disparity in academic performance between students from high and low socioeconomic backgrounds is equivalent to about three years of learning, a value similar to fifty years ago (Hanushek et al., 2020). Closing the achievement gap has been one of the top priorities of US policymakers over the past decades, but improvements have been modest despite the policies and resources deployed. Meanwhile, a growing body of research reports that interventions aimed at tailoring instruction to students' preparedness are particularly effective in fostering learning gains and reducing inequality, especially in early grades (see e.g., Banerjee et al., 2007; Kremer et al., 2013; Connor and Morrison, 2016; Connor et al., 2018). In practice, however, meeting the needs of academically diverse students could be a challenging task for educators, especially given the public good nature of instruction in a classroom environment (Lazear, 2001). A feasible alternative could then entail tailoring instruction to the needs of students with specific levels of ability.[1] Teachers can choose to orient instruction towards specific segments of the classroom based on the value that they attach to the achievement of students along the ability distribution, which can reflect a variety of factors, like personal preferences, incentives provided by education systems, or other institutional constraints.

Whether teachers value unequally the achievement of different students can determine how they adjust instruction based on the composition of the classroom. This, in turn, has potential implications for the distributional impact of student-classroom assignment policies, like the common practice of separating students into classrooms on the basis of ability (i.e., ability tracking). Besides, the extent to which teachers respond to the composition of the classroom can depend on potential match effects between the teacher's and the students' ability. Yet, teachers might also have incentives to stick to a predetermined curriculum or to simply follow specific teaching strategies regardless of the student composition in the classroom.

Existing empirical evidence seems to indicate the presence of specific patterns in teaching strategies across different countries. Recent studies show that educators in developing countries tend to direct their efforts towards better prepared students (e.g., Duflo et al., 2011; Gilligan et al., forthcoming; Cuesta et al., 2020), while teachers in US schools are more likely to target students in the lower part of the distribution as a result of incentive-based policies like No Child Left Behind (e.g., Reback, 2008; Neal and Schanzenbach, 2010; Deming and Figlio, 2016; Macartney et al., 2021). Yet, these findings are often inferred from student test scores, while evidence from direct information on teachers' instructional decision in the classroom is still scarce.

In this paper, I exploit unique data with rich information on teachers' instructional choices and

[1]The terms "ability", "readiness", "prior achievement", "prior knowledge", and "baseline knowledge" are used interchangeably throughout the paper.

teaching skills to address the following research questions: Do teachers assign unequal values the achievement of different students? Are teachers' instructional choices affected by the composition of the classroom? How does instruction shape the distribution of student achievement, and what are the implications for the impact of teacher-student assignment policies? I address these questions by developing and estimating an equilibrium model of endogenous instructional choices and student effort in a classroom environment. Teachers are endowed with teaching ability and choose the allocation of class time among different topics jointly with the amount of instructional effort to exert throughout the school year. Students, instead, are endowed with a level of baseline knowledge and choose how much learning effort to exert. Teacher and students strategically interact in a classroom environment, and their choices are modeled as the equilibrium of a static game of complete information. A technology of knowledge formation links instructional choices, teacher ability, and student inputs to the production of end-of-year knowledge, with the allocation of class time among different topics having a potentially heterogeneous impact on students with different levels of baseline knowledge. In particular, the parametric specification of the production function allows me to find the specific allocations of instructional time tailored to each student's level of prior knowledge. Moreover, the technology incorporates direct peer-to-peer spillovers not mediated by the teacher's behavior (e.g., originating from direct interactions among students). Availability of data on instructional choices and student characteristics allows me to empirically disentangle their separate contribution to the overall level of peer effects.

The model allows teachers to value differently the achievement of different students by attaching a specific weight to each student's end-of-year knowledge. Although not modeled in this framework, these weights can reflect the influence of various factors that potentially determine how teachers are rewarded for their students' performances, including monetary incentives or personal preferences. Moreover, these weights, combined with the specification of the knowledge production function, are able to generate the mechanism through which teachers orient their instruction towards students with specific levels of prior achievement. Specifically, teachers might optimally choose an allocation of class time closer to the one tailored to those students whose achievement is weighted the highest. Teachers also bear a cost of exerting effort and have preferences over time spent teaching specific topics and its alignment with the state-level curriculum standards. These standards represent the content and pace of instruction that, according to the educational authorities, teachers are supposed to follow in order to attain student proficiency by the end of the grade. Finally, students care about their end-of-year knowledge and also bear a learning effort cost.

The model is estimated using data from the Measures of Teaching Effectiveness (MET) project carried out by the Bill and Melinda Gates Foundation between 2009 and 2011. The dataset merges school administrative information on test scores and other student characteristics from five US public school districts with a large set of measures of teacher ability, teacher effort, student effort, and detailed information on class time allocation across topics.[2] In particular, the empirical analysis focuses on fourth grade math classrooms. The model is estimated through simulated maximum likelihood, which accounts for the potential presence of measurement error by exploiting the availability of a large constellation of measures of both student-level and teacher-level inputs.

The results from the estimated model suggest that teachers attach higher rewards to the achievement of students with lower levels of initial knowledge. These estimates turn out to be a good characterization of the incentives provided by the US education system, especially as documented by evidence on recent incentive-based policies like NCLB (see e.g. Macartney et al., 2021; Deming and Figlio, 2016). Furthermore, my estimates show that better prepared students tend to be more productive in learning new material, that teacher ability is positively related to students' learning, and that the allocation of class time across different topics has a significantly different impact on end-of-year knowledge depending on the student's level of preparation. As for direct peer-to-peer spillovers, the sign of the estimates uncover two main patterns. First, students in the lower and middle part of the distribution benefit from larger shares of higher-achieving peers, but only from adjacent quantiles. Second, only high-achieving students benefit from classmates with similar levels of prior knowledge.

Besides allowing for a more complete specification of knowledge accumulation process, the inputs included in the knowledge production function play a key role in controlling for factors potentially related to the non-random assignment of teachers to classrooms. Indeed, the latter is often considered a primary source of bias in the estimation of education production functions. To the extent that teacher assignments are based on prior test scores, teacher ability, or other observable characteristics, the inputs included are able to account for a wide range of potential confounding factors. Yet, assignment based on unobservables could still afflict the estimates. In order to check for the validity of the estimated model, I perform an out-of-sample validation exercise using data from the second year of the MET study in which teachers were randomly assigned to classrooms within each school. Specifically, I use the model estimates to predict measures of instruction and end-of-year knowledge in the second-year sample and then compare the simulated values with the actual data.[3] Results show that, although teachers were not randomly assigned in the sample used for the estimation, the model does a good job predicting second-year outcomes.

The estimated model allows me to run a counterfactual experiment where I implement ability tracking on 4^{th} grade math classes. To this end, I re-assign students to classrooms based on their prior test scores and simulate the outcomes under three alternative teacher assignment mechanisms: *(i)* random assignment, *(ii)* higher ability teachers to higher tracks and lower ability teachers to lower tracks (*positive assortative matching*), and *(iii)* higher ability teachers to lower tracks and lower ability

[2] There are a total of seven school districts participating in the MET study. However, two of these districts do not provide the necessary data to be included in the empirical analysis.

[3] Note that the simulated instructional time inputs cannot be compared to the actual data, as measures of class time allocation were not collected in the second year of the study.

teachers to higher tracks (*negative assortative matching*). I find that the effect of tracking on end-of-year knowledge depends significantly on the way teachers are assigned to classrooms. In particular, assigning better teachers to higher tracks can generate an increase in the achievement of students at the top of the distribution of about 0.16 standard deviations (SD), while performance of students in the middle and bottom terciles would decrease by 0.04SD. On the other hand, assigning the best teachers to lower-ability students increases the achievement of both high and low-achieving students by 0.05SD and 0.027SD, respectively. Repeating this counterfactual experiment assigning teachers to classrooms based on teaching experience yields results similar to mere random assignment. The main novelty of these findings is that, on top of accounting for teachers' behavioral response to changes in classroom composition, they shed light on a new dimension that determines the distributional effect of tracking, namely the teacher assignment mechanism.

These findings are partly driven by the way teachers adjust instruction to the specific track they are assigned to. In particular, teachers assigned to lower tracks tend to reallocate instructional time in a way that is better tailored to students' baseline knowledge, as well as to increase the amount of effort they exert. Both these responses are directly implied by the increased classroom homogeneity generated by the tracking policy. In fact, while tracking allows teachers to better match the pace of instruction to the students' initial knowledge, it also separates students whose achievement is more rewarding (i.e., weaker students) from those whose performance is less so. The model also allows me to disentangle the two peer effects channels operating under ability tracking. I find that ignoring the behavioral response of teachers to this policy can result into substantial bias in the estimated effects, often underestimating the potential benefits of tracking for students at the bottom of the distribution.

In a final analysis, I look at how curriculum standards impact the achievement of students at different levels of the distribution. Contrary to the beliefs of educational authorities, adhering to curriculum standards does not always translate into higher learning (see e.g., Polikoff and Porter, 2014). In fact, setting common standards has the potential drawback of curtailing flexibility, as they could impose a curriculum that is either too ambitious or too undemanding for the students in different schools and classrooms. To assess whether this is the case, I simulate a counterfactual experiment in which all teachers teach according to the state's curriculum standards. The results show that students along the entire distribution of prior knowledge would experience a decrease in end-of-year achievement.

The contribution of the present study spans several strands of the literature. First, this paper contributes to a recent literature on the impact of incentives on teacher rewards and student outcomes. Duflo et al. (2011) finds that the heterogeneous impact of tracking on achievement is consistent with the hypothesis that teachers in Kenyan schools tend to tailor their instruction to students at the top of the distribution. Using data on US schools, Macartney et al. (2021) find that the implementation of NCLB in North Carolina created a peak of test scores growth around proficiency cutoffs, while

Deming and Figlio (2016) report higher achievement gains of low-achieving students in schools that were more likely to be marked as "low performing" under an accountability program in Texas. In both cases, the authors interpret these results in terms of teachers directing their efforts towards students at the margin or in the lower tail of the achievement distribution in response to the incentive programs. A novelty of the present study is that it employs instructional choices data in order to investigate the extent to which teachers orient instruction towards specific groups of students, rather than inferring such behavioral responses from changes in distribution of student outcomes. My results are in line with prior evidence in that it confirms that teachers in US public schools attach higher rewards to achievement gains of students in lower quantiles.

Second, this paper contributes to the literature on the distributional impact of ability tracking on student outcomes. Evidence on the effect of tracking on student outcomes is still mixed. Fu and Mehta (2018) develop and estimate a model of endogenous tracking choices and parental investments, and find that tracking benefits only high-achieving students while being detrimental for those assigned to lower tracks. These results are consistent with early findings by Argys et al. (1996). Similar results are also found by Donaldson et al. (2017), who show that teachers assigned to lower tracks provide less emotional, organizational, and instructional support to students. On the other hand, Betts and Shkolnik (2000) find little differences in student outcomes between tracking and non-tracking schools, while Figlio and Page (2002) find that tracking might benefit low-ability students when accounting for endogenous sorting into schools. Duflo et al. (2011) use a randomized experiment to study the effect of two-way tracking in Kenyan schools. They find that tracking increases student achievement significantly across the entire ability distribution, and that these effects are likely driven by the behavioral response of teachers to the increased classroom homogeneity. My model expands on the theoretical and empirical findings of Duflo et al. (2011) by investigating how teachers' choices respond to tracking as well as how its impact on achievement might depend on the specific teacher-classroom assignment mechanism employed.

This paper also adds to a more general and well-established literature on peer effects in the classroom (e.g. Manski, 1993; Brock and Durlauf, 2001; Sacerdote, 2011, for a review). As pointed out by Sacerdote (2011), there are a large number of channels through which peers can affect student outcomes. In particular, both Duflo et al. (2011) and Todd and Wolpin (2018) highlight how peer spillovers can occur from the behavioral response of teachers to the distribution of student characteristics in the classroom. Similarly, Aucejo et al. (2021) employ data from the MET project and find that different teaching practices have different effects on student achievement depending on the composition of the classroom. Moreover, there is growing evidence that peer effects are non-linear and heterogeneous across students' own ability (e.g., Hoxby and Weingarth, 2005; Booij et al., 2017). The present paper contributes to this literature by explicitly modeling "indirect" peer effects stemming from teachers' response to classroom composition and allowing for (potentially non-linear) direct peer-to-peer spillovers. Finally, the present study contributes to a strand of research focused on the estimation of skills and education production functions (e.g., Ben-Porath, 1967; Todd and Wolpin, 2003, 2007; Cunha et al., 2010) as well as on the estimation of model of endogenous decisions by teachers and students in the classroom (e.g., Todd and Wolpin, 2018). This paper constitutes an addition to this literature by including endogenous instructional time allocation to the achievement production function, with its effect on learning gains being allowed to depend on the student's prior knowledge.

The rest of the paper is structured as follows. Section 2 describes the structure of the model and the specification of the knowledge formation technology. Section 3 analyses the identification of the model and discusses the estimation method; Section 4 describes the data and reports descriptive statistics of the final sample; Section 5 discusses the estimation results as well as internal and external validation of the model. Finally, Section 6 analyses the counterfactuals and policy experiments, whereas Section 7 contains concluding remarks.

2. A Model of Teacher's Instructional Decisions

This section presents an equilibrium model capturing the potential mechanisms underlying teacher instructional choices and peer effects. The model focuses on 4^{th} grade math classrooms. I assume that each teacher teaches only in one classroom. Each teacher chooses both teaching effort and class time allocation across topics given her preferences over the achievement of her students, over specific time allocations, and over the costs of exerting teaching effort and of deviating from curriculum standards. Students choose learning effort based on their preferences over their own achievement and based on the characteristics of their classmates. Teacher and students are assumed to make their choices simultaneously.[4]

2.1 Environment and Choices

Consider a teacher t teaching in a class composed by N_t students, each of them indexed by i. The teacher is endowed with a level of ability A_t, which affects the productivity of her instruction, and a total amount of class time over the entire school year, $\overline{\tau}_t$. The latter can be allocated among J different topics, where time spent on topic $j \in \{1, \ldots, J\}$ is denoted by $\tau_{tj} \in [0, \overline{\tau}_t]$. Define the class time allocation vector chosen by teacher t as $\boldsymbol{\tau}_t = (\tau_{t1}, \ldots, \tau_{tJ})$, with $\sum_{j=1}^{J} \tau_{tj} = \overline{\tau}_t$. On top of class time allocation, the teacher chooses the amount of instructional effort to exert in class, which is assumed to be a non-negative scalar e_t. Both $\boldsymbol{\tau}_t$ and e_t are assumed to be pure public inputs, thus excluding the possibility of individualized instruction or within-classroom ability grouping practices. Finally,

[4] The model is meant to be an approximation of the complex dynamics encompassing the interactions between teachers and students throughout the school year. For instance, one can envision a more complete dynamic model where teacher and students adjust their actions sequentially in a day-to-day basis.

each student i taught by teacher t starts with a level of initial knowledge in math, K_{0ti}, and exerts learning effort, $h_{ti} \geq 0$.

2.2 Knowledge Production Technology

Student i end-of-year knowledge in mathematics, K_{1ti}, is determined by the production function

$$K_{1ti} = \delta_0 K_{0ti} + \delta_1 K_{0ti}^{\gamma_0} A_t^{\gamma_1} e_t^{\gamma_2} h_{ti}^{\gamma_3} \prod_{j=1}^{J} \tau_{tj}^{\eta_{jq_i}} + \sum_{k=1}^{\overline{q}} \pi_{q_i k} f_{t,-i}^{k}, \tag{1}$$

where $\delta_0 K_{0ti}$ is the depreciated stock of prior knowledge, q_i denotes the $\overline{q}$-quantile of K_{0ti} (henceforth referred to as simply student i's quantile), and $\eta_{jq} \in (0,1)$ for each $j = 1,\ldots,J$. The last term in (1) captures potential "direct" peer-to-peer spillovers, with $f_{t,-i}^{k}$ the fraction of i's classmates in quantile k, and $\left(\pi_{qk}\right)_{k,q=1}^{\overline{q}}$ parameters to be estimated. This specification is flexible enough to allow for potential non-linearities in peer effects, where the distribution of peer characteristics can have a different effect on students at different quantiles of K_{0ti} (see e.g., Booij et al., 2017; Hoxby and Weingarth, 2005).[5] Equation (1) is assumed to satisfy constant returns to scale (CRS) in time inputs conditional on the quantile q, i.e., $\sum_{j=1}^{J} \eta_{jq} = 1$ for each $q = 1,\ldots,\overline{q}$. The specification in (1) follows the formulation of human capital production by Ben-Porath (1967) in that it posits a knowledge accumulation process where a flow of learning gains, or knowledge value-added (i.e., the second and third terms in the equation), is added to the level of existing stock of knowledge net of depreciation, $\delta_0 K_{0ti}$. The knowledge value-added captures the direct outcome of the learning process, where both instructional and non-instructional inputs are combined and transformed into additional knowledge.

The Cobb-Douglas specification of the first term of the knowledge value-added is consistent with the theoretical and empirical literature on learning and cognitive achievement. First, this functional form fits the intuitive idea that instruction is made of two complementary elements: 1) the content (i.e., what the teacher teaches), determined by the class time allocation $\boldsymbol{\tau}_t$, and 2) the delivery of such content to the students, here governed by instructional effort e_t.[6] In particular, the content of instruction includes both the curriculum (i.e., the set of topics the teacher allots positive amount of time to) as well as the pace, defined by the specific distribution of class time allocated among the topics covered in class (e.g., a slower pace of instruction could involve spending more time on basic topics and less on more advanced ones). Effort, instead, is meant to capture the degree to which the teacher takes actions aimed at delivering the content to the students in an effective way. Moreover, the model allows for the quality of instruction delivery, in terms of its impact on students learning gains, to depend on the ability of the teacher, A_t. The latter generally includes teacher skills like verbal ability

[5] An error term, modeled as classical measurement error, is going to be included in the empirical specification of the production function, as discussed in 3.1.2.

[6] Empirical evidence on the complementarity between content, delivery, and teacher ability can be found in Agodini and Harris (2014).

and content knowledge, which are considered among the most important attributes of teaching effectiveness.[7] Second, equation (1) implies that the time inputs in $\boldsymbol{\tau}_t$ (i.e. time spent in a each topic j, τ_{tj}) are all complement to each other. This is consistent with the literature on learning trajectories of students in different subjects (see e.g. Kilpatrick et al., 2001, 2003, for a review of of the theoretical and empirical studies of instruction and learning in mathematics). [8] Third, the elasticity parameters $(\eta_{jq})_{j=1}^{J}$ are quantile-specific, thus allowing time inputs to be more or less productive depending on the level of i's initial knowledge. As discussed below, this feature of the model allows teacher to tailor instruction to specific segments of the classroom. Finally, the complementarity of teacher inputs with K_{0ti} and h_{ti} is consistent with the idea that the effectiveness of instruction depends on the students' level of preparation and their level engagement in learning activities, where the latter can include time spent studying the subject, amount of attention during classes, or class disruption. The complementarity between instruction and student readiness is well-documented in the literature (see e.g., Bodovski and Farkas, 2007; Engel et al., 2013, 2016; Todd and Wolpin, 2018), while recent studies find both a positive impact of student effort on achievement (e.g. Burgess, 2016) as well an increase in the effectiveness of teacher effort when students are more engaged in learning inside and outside the classroom (see e.g. Todd and Wolpin, 2018).

An implication of the specification in (1) is that it allows me to find the optimal class time allocation for each student. Formally, this is given by $\widetilde{\boldsymbol{\tau}}_{ti} = \arg\max_{\boldsymbol{\tau}_t}(K_{ti})$ conditional on all other inputs and subject to the time constraints $\tau_{tj} \in [0, \overline{\tau}_t]$, $j = 1, \ldots, J$, and $\sum_{j=1}^{J} \tau_{tj} = \overline{\tau}_t$. Solving the maximization problem, we obtain

$$\widetilde{\boldsymbol{\tau}}_{ti} = (\overline{\tau}_t \eta_{1q_i} \ldots, \overline{\tau}_t \eta_{Jq_i}) = \overline{\tau}_t \boldsymbol{\eta}_{q_i} \equiv \widetilde{\boldsymbol{\tau}}_t^{q_i}, \tag{2}$$

where $\boldsymbol{\eta}_{q_i} = (\eta_{1q_i} \ldots, \eta_{Jq_i})$. Thus, $\widetilde{\boldsymbol{\tau}}_t^{q_i}$ represents the class time allocation *tailored* to all students at quantile $q = q_i$. This follows from a well-known property of the Cobb-Douglas with CRS and resources constraint, which implies that the optimal share of time allocated to each topic is given by the elasticity parameters $\boldsymbol{\eta}_q$. The tailored instruction $\widetilde{\boldsymbol{\tau}}_t^{q}$ is key in this model as it represents the channel through which teachers are able to target the instructional needs of specific students. In particular,

[7] The importance of teaching ability in explaining achievement gains is stressed by Kane et al. (2013), who show how a variety of research-based teaching effectiveness measures is able to strongly predict teachers' value-added. Moreover, the set of relevant attributes included in teacher ability can potentially go beyond content knowledge and verbal ability, as noted by Darling-Hammond and Youngs (2002) and Andrew et al. (2005).

[8] Complementarity arises naturally between many mathematical topics. For example, learning how to compute the area of a rectangle can reinforce the understanding of number multiplication. Yet, a potential drawback of the Cobb-Douglas specification in (1) is that it carries the strong assumption that $\tau_{tj} = 0$, for any j, implies zero learning gains. Although relevant from a theoretical point of view, this assumption has no particular implication in the specific application of the present study, as $\tau_{tj} = 0$ never occurs in the data. Moreover, this specification has the desirable feature of allowing the identification of the time allocation vectors tailored to each level of initial knowledge, as discussed below. A similar Cobb-Douglas specification for time inputs in the achievement production function has been also used by e.g., Del Boca et al. (2014).

teachers can orient instruction towards students at a specific quantile q by choosing a vector $\boldsymbol{\tau}_t$ closer to $\tilde{\boldsymbol{\tau}}_t^q$.

2.3 Curriculum Standards

The majority of the US state departments of education adopt curriculum standards, which are defined as the content students are supposed to know at the end of each grade and what teachers should teach in order to ensure students' proficiency. My model allows teachers to follow the state-level standards as defined by the vector $\boldsymbol{\varphi}_t = (\varphi_{t1}, \ldots, \varphi_{tJ})$, where each element φ_{tj} is defined as the amount of class time the teacher is supposed to spend on topic j. Notice that standards could be different across schools depending on the state they are located. This implies that if two teachers t and t' are located in the same state, then $\boldsymbol{\varphi}_t = \boldsymbol{\varphi}_{t'}$.

2.4 Preferences

Teachers have preferences over their students' knowledge as well as chosen instruction over the school year.[9] In particular, preferences are represented by the utility function

$$U_t = \sum_{i=1}^{N_t} \omega_{ti} K_{1ti} - \frac{e_t^2}{2} + \sum_{j=1}^{J} \left(\alpha_{1j} + \varepsilon_{tj}\right) \tau_{tj} - \sum_{j=1}^{J} \frac{\alpha_{2t}^j}{2} \left(\tau_{tj} - \varphi_{tj}\right)^2. \tag{3}$$

Equation (3) is composed by four terms. The first one represents preferences over students' end-of-year knowledge, specified as a weighted average of the elements in $\boldsymbol{K}_{1t} \equiv (K_{1ti}, \ldots, K_{1tN_t})$. The teacher attaches a (possibly different) value (or weight) to each student's knowledge level, which is captured by the student-specific parameter ω_{ti}. In particular, this weight is assumed to follow the parametric specification

$$\omega_{ti} = \sum_{q=1}^{\overline{q}} \mathbf{1}\{q = q_i\} \omega_1^q + W_{ti}' \boldsymbol{\omega}_2, \tag{4}$$

where the first term on the RHS captures the part of ω_{ti} determined by student i's baseline knowledge, which is represented by the quantile-specific parameter ω_1^q, and the last term allows ω_{ti} to depend on other students' or teacher's characteristics W_{ti}, like gender or race. The second term in (3) represents teacher's effort cost, which is assumed to be quadratic in e_t, whereas the third term captures teachers' preferences over the allocation of class time among classroom activities. Specifically, $(\alpha_{1k} + \varepsilon_{tk})$ is the marginal utility (cost) the teacher gets (bears) when increasing time spent on activity k (while holding $\boldsymbol{K}_{1t}$ fixed), with ε_{tk} a mean-zero preference shock and α_{1k} a parameter to be estimated. Finally, the last term represents teacher's utility (cost) of deviating from the state curriculum standards $\boldsymbol{\varphi}_t$. This

[9] Consistently with the literature on instructional effort choices, I assume that teachers do not account directly for students' future outcomes when making decisions (e.g. Barlevy and Neal, 2012; Macartney et al., 2021; Todd and Wolpin, 2018). In fact, the implicit assumption is that teachers care about students' future outcomes (like graduation, college enrollment, earnings etc.) only to the extent to which they are determined by knowledge produced during the school year they are teaching in.

term captures teacher t's preference for adhering to the standards. In particular, $\alpha_{2t}^{j} < 0$ implies a general compliance of the teacher with the standards on topic j, while a positive value indicates a willingness to depart from φ_{tj}. I model the parameter α_{2t}^{j} as

$$\alpha_{2t}^{j} = \alpha_{20j} + \alpha_{21}\phi_t,$$

where α_{20j}, $j = 1,\ldots,J$, and α_{31} are parameters to be estimated, and ϕ_t represents teacher t's preference over the alignment to the standards. Specifically, the latter captures both teachers' personal preferences as well as potential constraints imposed by the school or district.

Students have preferences over their end-of-year knowledge and their effort, as represented by the utility function

$$U_{ti} = \psi_{ti} K_{1ti} - \frac{h_{ti}^2}{2}, \tag{5}$$

where ψ_{ti} captures student-specific preference over her end-of-year knowledge.

2.5 Model Solution and Equilibrium

Both the teacher and the student's reaction functions are obtained through the maximization of their utility conditional on the information available. I assume that all teacher and students characteristics are known to the players when they make the decisions. That is, $G_t \equiv \left(A_t, \varphi_t, \boldsymbol{\varepsilon}_t, \boldsymbol{\phi}_t, (K_{0ti}, \psi_{ti})_{i=1}^{N_t}\right)$ is common knowledge among the teacher and students in the classroom. For given G_t and the level of effort exerted by students in the classroom, $\boldsymbol{h}_t \equiv (h_{t1},\ldots,h_{tN_t})$, the teacher chooses effort e_t and class time allocation $\boldsymbol{\tau}_t$ in order to maximize (3) subject to the choice variables constraints. Formally,

$$\max_{e_t,\boldsymbol{\tau}_t} \left[\sum_{i=1}^{N_t} \omega_{ti} K_{1ti} - \frac{e_t^2}{2} + \sum_{j=1}^{J} \left(\alpha_{1j} + \varepsilon_{tj}\right)\tau_{tj} - \sum_{j=1}^{J} \frac{\alpha_{2t}^{j}}{2} \left(\tau_{tj} - \varphi_{tj}\right)^2 \right]$$

$$s.t. \quad e_t \geq 0, \quad \tau_{tj} \in [0, \overline{\tau}_t], \text{ for } j = 1,\ldots,J, \quad \sum_{j=1}^{J} \tau_{tj} = \overline{\tau}_t. \tag{6}$$

Substituting for $\tau_{tJ} = \overline{\tau}_t - \sum_{j=1}^{J-1} \tau_{tj}$ and taking the first-order conditions, we obtain the following equations for an interior solution for the reaction functions of effort, $e_t^*(\boldsymbol{h}_t)$, and time spent on each activity $k = 1\ldots, J-1$, $\tau^*_{tk}(\boldsymbol{h}_t)$,

$$\gamma_2 e_t^{\gamma_2 - 1} \sum_{i=1}^{N_t} \omega_{ti} \delta_1 K_{0ti}^{\gamma_0} A_t^{\gamma_1} h_{ti}^{\gamma_3} \prod_{j=1}^{J} \tau_{tj}^{\eta_{jq}} - e_t = 0 \tag{7a}$$

$$\sum_{i=1}^{N_t} \omega_{ti} \delta_1 K_{0ti}^{\gamma_0} A_t^{\gamma_1} e_t^{\gamma_2} h_{ti}^{\gamma_3} (\eta_{kq_i} \tau_{tk}^{-1} - \eta_{Jq} \tau_{tJ}^{-1}) \prod_{j=1}^{J} \tau_{tj}^{\eta_{jq}} +$$
$$+ (\widetilde{\alpha}_{1k} + \widetilde{\varepsilon}_{tk}) - \alpha_{2t}^k (\tau_{tk} - \varphi_{tk}) + \alpha_{2t}^J (\tau_{tJ} - \varphi_{tJ}) = 0, \tag{7b}$$

where $\widetilde{\alpha}_{1j} \equiv (\alpha_{1j} - \alpha_{1J})$ and $\widetilde{\varepsilon}_{1j} \equiv (\varepsilon_{1j} - \varepsilon_{1J})$. Equation (7a) represents the optimality condition for e_t, which equates the marginal utility of students' end-of-year knowledge from a change in e_t with the marginal cost of effort. Notice that, given $e_t \geq 0$, equation (7a) holds as long as the weighted average in the first term is positive. Otherwise, the teacher optimally chooses to exert no effort by setting $e_t^* = 0$. This equation determines the relationship between instructional effort and the other inputs. In particular, it shows that teachers respond to the classroom distribution of initial knowledge, $\boldsymbol{K}_{0t}$, and how this relationship is governed by the interaction of the values attached to each student end-of-year knowledge, $(\omega_{sti})_{i=1}^{N_t}$, with the the other inputs determining the productivity of effort. Indeed, the more productive is effort in producing knowledge, the higher is the value of e_t^* the teacher chooses. These implications also characterize the relationship between instructional effort and class time allocation given the results obtained in 2.2. In fact, the closer is time allocation $\boldsymbol{\tau}_t$ to the value tailored to the students whose achievement teacher t finds most rewarding, the more effort she will exert.[10] The optimality condition for an interior solution of each time input τ_{tk}, $k = 1 \cdots, J-1$, is represented by equation (7b). Although the way $\boldsymbol{\tau}_t^*$ is related to the value of other inputs and parameters is more complicated compared to the one with effort, the interaction between the weights $(\omega_{ti})_{i=1}^{N_t}$ and the other production function parameters governs also the relationship between class time allocation and the composition of the classroom. Yet, probably the most important aspect of these FOCs is the presence of $\boldsymbol{h_t}$ and $\boldsymbol{K}_{0t}$ in both (7a) and (7b), which determines how peer effects operate through the teacher's instruction (i.e., the *indirect* channel of peer effects). These equations also show how the sign and magnitude of such peer effects are non-trivial. For instance, even assuming $\gamma_0, \gamma_3 > 0$, higher levels of initial knowledge and student effort do not guarantee an increase in instructional effort, especially in case the teacher attaches a higher weight to the achievement of students in lower quantiles.

Given the effort levels exerted by her classmates, $\boldsymbol{h}_{t,-i}$, and the teacher's instruction, e_t and $\boldsymbol{\tau}_t$,

[10] Formally, if we define $q' \in \arg\max_q \{\omega_1^q\}_{q=1}^Q$ (i.e. the quantile of those students whose achievement the teacher attaches the highest value), the lower is the distance between $\boldsymbol{\tau}_t$ and $\widetilde{\boldsymbol{\tau}}_t^{q'}$ (from (2)), the higher will be $\prod_{j=1}^J \tau_{tj}^{\eta_{jq}}$ and, in turn, e_t^*.

student i chooses h_{ti} in order to maximize her utility,

$$\max_{h_{ti}} \left[\psi_{ti} K_{1ti} - \frac{h_{ti}^2}{2} \right] \quad s.t. \quad h_{ti} \geq 0, \tag{8}$$

The FOC for an interior solution of the effort reaction function $h_{ti}(\boldsymbol{h}_{t,-i}, e_t, \boldsymbol{\tau}_t)$ is then

$$\gamma_3 \psi_{ti} h_{ti}^{\gamma_3 - 1} \delta_1 K_{0ti}^{\gamma_0} A_t^{\gamma_1} e_t^{\gamma_2} \prod_{j=1}^{J} \tau_{tj}^{\eta_{jq}} - h_{ti} = 0. \tag{9}$$

The positive effort Nash equilibrium is given by the solution to the system of equations composed by the reaction functions determined by (7a), (7b), and (9) (for all $i = 1, \ldots, N_t$). To ease notation, define $D_{ti} \equiv \delta_1 K_{0ti}^{\gamma_0} A_t^{\gamma_1}$ and $F_q(\boldsymbol{\tau}_t) \equiv \prod_{j=1}^{J} \tau_{tj}^{\eta_{jq}}$, and also $\widetilde{\alpha}_{1k} \equiv \alpha_{1k} - \alpha_{1J}$ and $\widetilde{\varepsilon}_{tk} \equiv \varepsilon_{tk} - \varepsilon_{tJ}$. The time constraint implies $\tau_{tJ} = \overline{\tau}_t - \sum_{j=1}^{J-1} \tau_{tj}$. The positive effort equilibrium profile $\{e_t^*, \boldsymbol{\tau}_t^*, (h_{ti}^*)_{i=1}^{N_t}\}$ satisfies:

$$e_t^* = \left[\gamma_2 (\gamma_3)^{\frac{\gamma_3}{2-\gamma_3}} \sum_{\ell=1}^{N_t} \omega_{t\ell} [D_{t\ell} F_{q_\ell}(\boldsymbol{\tau}_t^*)]^{\frac{2}{2-\gamma_3}} (\psi_{t\ell})^{\frac{\gamma_3}{2-\gamma_3}} \right]^{\frac{2-\gamma_3}{2(2-\gamma_2-\gamma_3)}}$$

$$h_{ti}^* = \left[\psi_{ti} D_{ti} F_q(\boldsymbol{\tau}_t^*) \right]^{\frac{1}{2-\gamma_3}} \left[\gamma_2 (\gamma_3)^{\frac{2-\gamma_2}{\gamma_2}} \sum_{\ell=1}^{N_t} \omega_{t\ell} [D_{t\ell} F_{q_\ell}(\boldsymbol{\tau}_t^*)]^{\frac{2}{2-\gamma_3}} (\psi_{t\ell})^{\frac{\gamma_3}{2-\gamma_3}} \right]^{\frac{\gamma_2}{2(2-\gamma_2-\gamma_3)}}$$

and, for $k = 1, \ldots, J-1$,

$$\sum_{i=1}^{N_t} \omega_{ti} D_{ti} e_{st}^{*\gamma_2} h_{ti}^{*\gamma_3} (\eta_{kq_i} \tau_{tk}^{*-1} - \eta_{Jq} \tau_{tJ}^{*-1}) F_q(\boldsymbol{\tau}_t^*) + (\widetilde{\alpha}_{1k} + \widetilde{\varepsilon}_{tk}) -$$
$$- (\alpha_{20k} + \alpha_{21} \phi_t) \left(\tau_{tk}^* - \varphi_{tk} \right) + (\alpha_{20J} + \alpha_{21} \phi_t) \left(\tau_{tJ}^* - \varphi_{tJ} \right) = 0.$$

3. Estimation

In the empirical specification of the model, both inputs and outputs are assumed to be latent factors measured with error. The factor model allows me to correct for variables mis-measurement and the arbitrariness of their scales. This structure is in line with recent literature on child skills development (e.g. Cunha et al., 2010; Agostinelli and Wiswall, 2016) and similar to the specification employed by Todd and Wolpin (2018). This section describes the structure imposed to the latent factors as well as the system of measurement equations.

3.1 Latent Factors and Measurement Structure

3.1.1 Latent Factors Structure of Endowments

Each exogenously determined latent input $\theta \in \{A, \phi, K_0, \psi\}$ is assumed to depend linearly on a vector of initial conditions X^θ and on one or more random effects. Formally, the teacher's ability A_t, and preference over curriculum standards adherence ϕ_t are specified as

$$log(A_t) = X_t^A \boldsymbol{\beta}^A + \upsilon_t^A, \tag{10a}$$

$$log(\phi_t) = X_t^\phi \boldsymbol{\beta}^\phi + \upsilon_t^\phi \tag{10b}$$

with υ_t^A and υ_t^ϕ representing teacher-level unobserved error terms. Similarly, student i's baseline knowledge K_{0ti} and preference over own end-of-year knowledge ψ_{ti} are specified as

$$K_{0ti} = X_{ti}^{K_0} \boldsymbol{\beta}^{K_0} + \upsilon_t^{K_0} + \zeta_{ti}^{K_0} \tag{11a}$$

$$log(\psi_{ti}) = X_{ti}^\psi \boldsymbol{\beta}^\psi + \upsilon_t^\psi + \zeta_{ti}^\psi, \tag{11b}$$

where $\upsilon_t^{K_0}$ and υ_t^ψ, and $\zeta_{ti}^{K_0}$ and ζ_t^ψ, are are teacher-level and student-level unobserved components, respectively.[11] The log-linear specification of A_t, ϕ_t, and ψ_{ti} guarantees that these inputs take only positive values. Equation (11a) defines K_{0ti} as a linear function of the exogenous determinants $X_{ti}^{K_0}$.[12] The error terms at each separate level are allowed to be correlated across factors and are assumed to be orthogonal to the exogenous variables $X_{ti} \equiv (\boldsymbol{\varphi}_t, X_t^A, X_t^\phi, X_{ti}^{K_0}, X_{ti}^\psi)$, to each other, and to be mean zero and jointly normally distributed. Formally, $\boldsymbol{\upsilon}_t | X_{ti} \sim N(\mathbf{0}, \Sigma_\upsilon)$, $\boldsymbol{\zeta}_{ti} | X_{ti} \sim N(\mathbf{0}, \Sigma_\zeta)$, and $\boldsymbol{\upsilon}_t \perp \boldsymbol{\zeta}_{ti}$, where $\boldsymbol{\upsilon}_t \equiv (\upsilon_t^A, \upsilon_t^\phi, \upsilon_t^{K_0}, \upsilon_t^\psi)$ and $\boldsymbol{\zeta}_{ti} \equiv (\zeta_{ti}^{K_0}, \zeta_{ti}^\psi)$. Finally, the latent factors of teacher effort e_t, time allocation $\boldsymbol{\tau}_t$, student effort h_{ti}, and end-of-year knowledge K_{1ti} are endogenously determined by the equations (7a), (7b), (9), and (1), respectively.

3.1.2 Measurement Equations Structure

Both the endowments A_t, ϕ_t, and ψ_{ti}, and the endogenous variables e_t and h_{ti} are assumed to be latent factors measured with error. Dropping the subscripts to simplify the notation, let M_θ be the number of distinct measures and $Z^{\theta m}$ be the m-th measure latent factor $\theta \in \{A, \phi, \psi, K_1, e, h\}$, respectively. Each measure $Z^{\theta m}$ is allowed to be either continuous or ordinal. In particular, define

$$Z^{\theta m*} = \mu_0^{\theta m} + \mu_1^{\theta m} log(\theta) + \varsigma^{\theta m}, \qquad \text{for } \theta \in \{A, \phi, \psi\}, \quad m = 1, \ldots, M_\theta,$$

$$Z^{\theta m*} = \mu_0^{\theta m} + \mu_1^{\theta m} \theta + \varsigma^{\theta m}, \qquad \text{for } \theta \in \{K_1, e, h\}, \quad m = 1, \ldots, M_\theta.$$

Continuous measures are then defined as $Z^{\theta m} = Z^{\theta m*}$, while ordinal measures are defined as step functions with the latent variable equal to $Z^{\theta m*}$. Both baseline knowledge, K_{0ti}, and time inputs, $\boldsymbol{\tau}_t$, are assumed to be measured without error. Finally, I assume classical measurement errors together with joint normality, that is $\boldsymbol{\varsigma}_{ti} \equiv (\varsigma_t^{A,m}, \varsigma_{ti}^{K_1,m}, \varsigma_{ti}^{h,m}, \varsigma_t^{e,m},) \sim \mathcal{N}(\mathbf{0}, \Sigma_\varsigma)$,, where Σ_ς is a diagonal

[11] The model could also allow for school-level random effects. However, observations entailing only one teacher per school are quite frequent in the sample used for the estimation. As a result, separately identifying teacher and school-level effects would be a demanding task.

[12] As discussed in the next sections, log-linearity is not applied to K_{0ti} for compatibility with the measures of K_{1ti}.

variance-covariance matrix and $\boldsymbol{\varsigma}_{ti}$ with assumed orthogonal to all the observed and unobserved components of the latent factors.

3.1.3 Further Assumptions and Discussion

In order to bring the model to the data, it is necessary to first discuss some issues related to the measures available as well as to some necessary restriction to be imposed to the model. A first issue is given by the information available on instructional time inputs, as the MET data does not provide variables on $\boldsymbol{\tau}_t$ expressed in terms of time (e.g. hours, days, or weeks). Instead, data on class time allocation is available only in terms of fractions of total class time, $\boldsymbol{\tau}_t/\overline{\tau}_d$. In order to mitigate the potential consequences from the lack of information on $\overline{\tau}_t$, I allow the parameter δ_1 to be district-specific.[13] This assumption seems particularly suited to the data, as there is evidence that schools participating in the MET study have to abide to a specific total number of school days and class hours determined by the school district (with only few exceptions). This implies that total class time $\overline{\tau}_t$ is going vary for the most part between and not within districts.[14] As for curriculum standards, Section 2.3 points out that states do not actually provide the time variables $\boldsymbol{\varphi}_t$, but rather some documents which detail what skills a typical student is supposed to acquire in each subject by the end of each grade. Given that exact data on $\boldsymbol{\varphi}_t$ is not available, I will use information on the state test content collected by the MET study as a proxy of the standards.[15] Test content variables are also expressed as fractions, thus making them comparable to the curriculum data discussed above. The idea is that, to the extent that the state test is aimed at measuring students' proficiency, its content should reflect the educational standards set by the state. Moreover, there is evidence of alignment between tests and standards content, as shown by Polikoff et al. (2011).

For the empirical specification of the weights $(\omega_1^q)_{q=1}^{\overline{q}}$ and the elasticities $(\eta_{jq})_{j=1}^{J})_{q=1}^{\overline{q}}$, I use terciles, i.e. $\overline{q} = 3$. An additional specification issue concerns the variables to include as determinants of teacher's preferences, W_{ti}, in equation (4). For this I follow recent empirical evidence on gender stereotypes (Carlana, 2019) and ethnicity role model effects (Gershenson et al., 2018) and include both teacher-level and student-level gender and race dummies. A third issue entails the inclusion of class size effects on K_{1ti}. To account for that, I follow Todd and Wolpin (2018) and allow the elasticity of effort to depend on N_t through the equation $\gamma_2 = \gamma_{20} + \gamma_{21} N_t$. Moreover, in order to ensure a solution for optimal teacher and student effort, e_t^* and h_{ti}^*, I assume $\gamma_2, \gamma_3 \in (0,2)$. The parametric

[13]Indeed, a large body of research shows that total instructional time has a significant impact on student achievement. For a recent review of the literature, see Gromada and Shewbridge (2016).

[14]If total class time were indeed fixed to $\overline{\tau}_d$ across all schools within each district d, we can obtain the empirical specification of the knowledge production function by dividing and multiplying the knowledge value-added in equation (1) by $\overline{\tau}_d$, redefine each time input in fractional terms, $\tau_{tj}/\overline{\tau}_t$, for $j = 1,\ldots,J$, and define the district-specific parameter (to be estimated) as $\delta_{1d} \equiv \delta_1 \overline{\tau}_d$.

[15]The MET study actually provides data on the state standards content, whose variables are measured as fractions of "items" in the curriculum standards document about each single topic. As it is not clear whether the fraction of items in such documents is a good measure of the "weight" a state gives to each topic, using test content data seems a better option.

specification is then completed by imposing distributional assumptions on the the preference shocks $\widetilde{\boldsymbol{\varepsilon}}_t = (\widetilde{\varepsilon}_{t1}, \ldots, \widetilde{\varepsilon}_{tJ-1})$, which are assumed to be jointly normally distributed with mean zero, covariance matrix $\Sigma_{\widetilde{\varepsilon}}$, and orthogonal to the random effects $(\boldsymbol{\upsilon}_t, \boldsymbol{\zeta}_{ti})$ and measurement errors $\boldsymbol{\varsigma}_{ti}$.

A final concern is about the possibility of corner solutions in either exerted effort ($e_t^* = 0$ and h_{ti}^*) or in the time allocation choice ($\tau_{tk} = 0$ for some $k = 1\ldots, J-1$). In fact, a complete description of the model would necessitate an analysis of the conditions on the parameters, latent factors, and preference shocks values that give rise to each specific corner solution. However, since none of the measures of instruction used in my analysis are consistent with corner solutions, the empirical specification employs the FOCs in (7a), (7b), and (9) as the only conditions of optimality required to estimate the model.

3.2 Identification

To illustrate the sources of identification for the knowledge technology and preferences parameters, I first analyze the case of no measurement error. With perfect measures, the knowledge production function parameters $(\delta_0, \delta_1, \gamma_0, \gamma_1, \gamma_{20}, \gamma_{21}, \gamma_3, ((\eta_{jq})_{j=1}^J)_{q=1}^3, (\pi_{q1}, \pi_{q3})_{q=1}^3)$ are identified through independent variation in the observable inputs $(K_{0ti}, A_t, e_t, h_{ti}, \boldsymbol{\tau}_t)$ and end-of-year knowledge K_{1ti}, upon the normalization of $\pi_{q2} = 0$ for $q = 1, 2, 3$. As for the utility function parameters, the main source of identification comes from data on instructional choices together with variation in classroom composition characteristics. Specifically, the weights parameters $((\omega_1^q)_{q=1}^3, \boldsymbol{\omega}_2)$ are identified off variations in $(\boldsymbol{K}_{0t}, \boldsymbol{W}_t)$, time allocation choices, $\boldsymbol{\tau}_t$, and student and teacher effort, h_{ti} and e_t. Finally, the utility parameters $(\alpha_{1j}, \alpha_{2j})_{j=1}^J$ and the preference shocks covariance matrix $\Sigma_{\widetilde{\varepsilon}}$ are identified from the distributional moments of observed time inputs $\boldsymbol{\tau}_t$ and curriculum standards $\boldsymbol{\varphi}_t$ combined with the equations in (7b).

Turning to the latent factors model considered in this paper, the identification argument follows the one in Todd and Wolpin (2018). In particular, let the optimal instructional choices from (7a)-(7b) and the end-of-year knowledge production function (1) be represented as functions of the exogenous initial conditions, X_{ti}, and the random shocks $(\boldsymbol{\upsilon}_t, \boldsymbol{\zeta}_{ti}, \widetilde{\boldsymbol{\varepsilon}}_t)$. Formally

$$e_t^* = a_e(X_t, \boldsymbol{\upsilon}_t, \boldsymbol{\zeta}_t, \widetilde{\boldsymbol{\varepsilon}}_t) \tag{12a}$$

$$\tau_{tj}^* = a_{\tau_j}(X_t, \boldsymbol{\upsilon}_t, \boldsymbol{\zeta}_t, \widetilde{\boldsymbol{\varepsilon}}_t), \qquad j = 1, \ldots, J-1 \tag{12b}$$

$$h_{ti}^* = a_h(X_{ti}, \boldsymbol{\upsilon}_t, \boldsymbol{\zeta}_{ti}, \widetilde{\boldsymbol{\varepsilon}}_t) \tag{12c}$$

$$K_{1ti} = a_{K_1}(X_{ti}, \boldsymbol{\upsilon}_t, \boldsymbol{\zeta}_{ti}, \widetilde{\boldsymbol{\varepsilon}}_t), \tag{12d}$$

where $X_t = (X_{t1}, \ldots, X_{tN_t})$ and $\boldsymbol{\zeta}_t = (\boldsymbol{\zeta}_{t1}, \ldots, \boldsymbol{\zeta}_{tN_t})$. Consider now a system of equations that combines: *(i)* the exogenous latent factor equations from (10a), (10b), and (11b) with the measurement equations; *(ii)* the equations on endogenous effort, (12a), (12c) with their measurements; and *(iii)* (11a) and (12b), for $j = 1,\dots,J$, assumed to be measured without error. This system is a measurement model for the latent factors $(\boldsymbol{\upsilon}_t, \boldsymbol{\zeta}_{ti}, \widetilde{\boldsymbol{\varepsilon}}_t)$ analogous to (3.7) in Cunha et al. (2010). As a result, it is possible to invoke Theorem 2 from the same paper in order to identify both utility and production function parameters.

Parameters of the exogenous latent factors equations (10a), (10b), (11b), and (11a), and of the measurements equations are identified through observable determinants X_{ti} and multiple measurements of each latent variable (upon necessary normalizations). Identification of the latent and measurement equations for the exogenous inputs $(A_t, \phi_t, \psi_{ti}, K_{0ti})$ follows the canonical arguments of structural equation modeling as in Goldberger (1972). Consider a latent factor $\theta \in \{A, \psi, \phi, K_0\}$, with subscripts dropped whenever it is not confusing to do so. First, I normalize the intercept and slope of measure $m = 1$ (without loss of generality) by setting $\mu_0^{\theta 1} = 0$ and $\mu_1^{\theta 1} = 1$. Given the orthogonality assumptions of both unobserved random components and measurement errors, the latent equation parameters are identified by regressing the first measure $Z^{\theta 1}$ on X^{θ}, that is

$$\boldsymbol{\beta}^{\theta} = E[(X^{\theta})' X^{\theta}]^{-1} E[(X^{\theta})' Z^{\theta 1}].$$

Once the parameters $\boldsymbol{\beta}^{\theta}$ are known, the slopes and intercepts of the remaining measurement equations $m = 2,\dots,M_{\theta}$ for $\theta \in \{A, \psi, \phi\}$ are identified as follows. First, regress each measurement $Z^{\theta m}$ on X^{θ} and obtain $\widetilde{\boldsymbol{\mu}}^{\theta m} = E[(X^{\theta})' X^{\theta}]^{-1} E[(X^{\theta})' Z^{\theta m}]$, and then compute

$$\mu_1^{\theta m} = \widetilde{\mu}_j^{\theta m} / \beta_j^{\theta}, \qquad \mu_0^{\theta m} = \widetilde{\mu}_0^{\theta m} - \mu_1^{\theta m} \beta_0^{\theta},$$

for an arbitrary j^{th} element of $\widetilde{\boldsymbol{\mu}}^{\theta m}$, $j \geq 2$, and with β_0^{θ} being the latent factor equation constant. It is then possible to pin down Σ_{υ}, and Σ_{ζ} by computing the covariances between measures. In particular, the diagonal elements $\sigma^2_{\upsilon\theta}$, and $\sigma^2_{\zeta\theta}$ are obtained by

$$\sigma^2_{\zeta\theta} = Cov(Z_{ti}^{\theta 1}, Z_{ti}^{\theta m}) / \mu_1^{\theta m} - Var(X_{ti}^{\theta} \boldsymbol{\beta}^{\theta}) - \sigma^2_{\upsilon\theta},$$

where $m \geq 2$ for $\theta \in \{A, \psi, \phi\}$ and $m = 1$ for $\theta = K_0$, $(\overline{Z}_t^{\theta 1}, \overline{X}_t^{\theta})$ are class-level means, and the last equation holds only for the student-level latent factor, ψ. Similarly, the off-diagonal elements $\sigma_{\upsilon\theta\theta'}$, and $\sigma_{\zeta\theta\theta'}$, for $\theta, \theta' \in \{A, \psi, \phi, K_0\}$, $\theta \neq \theta'$, are determined as

$$\begin{aligned}
\sigma_{\upsilon\theta\theta'} &= Cov(\overline{Z}_t^{\theta 1}, \overline{Z}_t^{\theta' 1}) - Cov(\overline{X}_t^{\theta} \boldsymbol{\beta}^{\theta}, \overline{X}_t^{\theta'} \boldsymbol{\beta}^{\theta'}) \\
\sigma_{\zeta\theta\theta'} &= Cov(Z_{ti}^{\theta 1}, Z_{ti}^{\theta' 1}) - Cov(X_{ti}^{\theta} \boldsymbol{\beta}^{\theta}, X_{ti}^{\theta'} \boldsymbol{\beta}^{\theta'}) - \sigma_{\upsilon\theta\theta'}.
\end{aligned}$$

As a last step, the measurement error variances for teacher ability and student inputs measures are obtained as

$$\sigma^2_{\varsigma Am} = Var(Z_t^{Am}) - Var(X_t^A \boldsymbol{\beta}^A) - \sigma^2_{\upsilon A} \qquad m = 1, \ldots, M_A$$
$$\sigma^2_{\varsigma hm} = Var(Z_{ti}^{hm}) - Var(X_{ti}^h \boldsymbol{\beta}^h) - \sigma^2_{\upsilon h} - \sigma^2_{\zeta h}, \qquad m = 1, \ldots, M_h.$$

Finally, given that all the production function parameters are identified, the measurement equations parameters related to the latent student and teacher effort, $(\mu_0^{hm}, \mu_1^{hm})_{m=1}^{M_h}$ and $(\mu_0^{em}, \mu_1^{em})_{m=1}^{M_e}$, are identified from the multiple effort measures available in the data.

3.3 Likelihood Function

Estimation is carried out through simulated maximum likelihood (SML). Let Θ be the vector of all the parameters of the model to be estimated (including production function, preferences, exogenous latent factors, and measurement equations). The likelihood contribution of teacher/classroom t given Θ is represented by the joint density of the measurements of the latent factors for the teacher and all N_t students, denoted by $\mathscr{M}_t$, conditional on the initial conditions $\mathscr{X}_t \equiv (X_{t1}, \ldots, X_{tN_t}, \boldsymbol{\varphi}_t)$. Denoting the simulated likelihood contribution of teacher t by $L_t(\Theta|\mathscr{M}_t; \mathscr{X}_t)$, the likelihood function to be maximized over Θ is

$$\mathscr{L}(\Theta) = \prod_{t=1}^{T} L_t(\Theta|\mathscr{M}_t; \mathscr{X}_t),$$

where T is the total number of classrooms/teachers in the sample. A complete description of likelihood function formulas and of the estimation procedure is reported in the Appendix.

4. Data

For the empirical part of this paper I use data from the Measurements of Effective Teaching (MET) project, which was run by the Bill and Melinda Gates between 2009 and 2011. This study was conducted in two years in seven large US school districts and involved 2714 4^{th}-to-9^{th} grade Math and English Language Arts (ELA) teachers in 317 schools.[16] The main goal of this project was to assess the ability of a large set of research-based indicators of teacher quality to identify effective teachers. Moreover, in order to ensure validity of the estimates, teachers in the second year of the study were randomly assigned to classrooms within each school (while in the first year the assignment was performed as usual). The data collected by the MET study include detailed information on: *i)* teaching practices in the classroom from both video-recorded lessons and surveys taken by both teachers and students; *ii)* topics covered in end-of-grade state tests; *iii)* self-reported student information on own effort and home environment; *iv)* end-of-grade state tests scores in various subjects and teacher and

[16] The seven school districts included in the MET study are: Charlotte-Mecklenbourg Schools (NC), Dallas Independent School District (TX), Denver Public Schools (CO), Hillsborough County Public Schools (FL), Memphis City Schools (TN), the New York City Department of Education (NY), and the Pittsburgh Pubilc Schools (PA).

student demographics from administrative district data. All these variables provide information used to measure the latent inputs and outputs of the model. The estimation is carried out using a subsample of the first-year MET data including 4^{th} grade Math teachers who took the Survey of Enacted Curriculum (SEC).[17] Originally developed by Porter and Smithson (2001) to study the alignment of classroom instruction with curriculum standards and test content, this survey asks teachers to report their class time allocation throughout the school year across an exceptionally fine-grained array of different topics spanning all school grades. These answers were then converted and re-expressed as fractions of total class time.[18] The survey was conducted only in the first year of the MET study, thus not allowing me to exploit the second-year random assignment of teachers to estimate the model. Nevertheless, second-year data is used to perform an out-of-sample validation exercise (see Section 5.3). The final sample includes 101 teachers and 2532 students from 85 schools in 5 school districts.[19]

Table 1 displays descriptive statistics of the student and teacher characteristics determining student effort and teacher ability. Panel A shows that students in the sample are on average 9-10 years old (as expected for 4^{th} graders) and the gender ratio among them is almost 1:1, with a slightly higher percentage of females. The majority of the student population is black (43%) followed by white (26%) and Hispanic students (25%). About 6% of the students have been identified as gifted, while almost 9% are placed in special education programs (SpEd) due to learning disabilities and 17% are labeled as English language learners (ELL). As regards to students' socioeconomic status (SES) and family/home environment, almost half of the students in the estimating sample receive either free or reduced-price lunch and 88% of them has at least one computer at home. Finally, nearly half of the responding students possess at least 25 books in their bedroom, 40% of the students report to have always a quite place to study at home, and 69% has always a person at home who can help with homework. The sample is clearly not representative of the US student population, as it includes a much higher percentage of black and Hispanic students and a slightly lower percentage of students served by SpEd programs compared to national averages. As for teacher characteristics, Panel B shows summary statistics of the determinants of teacher ability. The majority of the teachers (83%) are generalist, i.e. teach both Math and ELA to the same classroom. As for characteristics related to their human capital, teachers in the sample have on average 6.5 years of teaching experience in the school district with a quite significant variability, and about half of the teachers possess a Master's degree.

[Table 1 here]

Table 2 reports descriptive statistics on class time allocation and curriculum state standards. Top-

[17] The full sample of teachers taking the SEC included 4^{th} and 8^{th} grade Math and ELA teachers.

[18] As pointed out in Section 3.1.3, the MET data does not provide teachers' original answers on the actual time spent on different topics. As a reference point, Trends in International Mathematics and Science Study (TIMSS) reports that in 2015 teachers in the US spent on average 216 class hours.

[19] Unfortunately, confidentiality restrictions do not allow to disclose the exact identity of these five districts out of the seven listed above.

ics are aggregated in five different groups representing common areas of mathematics covered in 4^{th} grade classes, namely: place value, rounding, addition, and subtraction (group 1); multi-digit multiplication and division (group 2); shapes, angles, and geometry (group 3); fractions and decimals (group 4); unit conversion and measurement (group 5).[20] On average, teachers split 3/4 of the school year evenly between teaching multi-digit multiplication and division (26%) fractions and decimals (25.5%) and unit conversion and measurement (24.2%). The remaining time is then largely devoted to geometry topics (17.5%), while only a smaller fraction of class time focuses on more basic topics like place value, rounding, addition, and subtraction of whole numbers (6.8%). Columns 2 to 5 show the variation of these time allocations across teachers. There is a 2 percentage points variation between the first and third tercile in the percentage of class time devoted to place value, rounding, addition, and subtraction, whereas the difference is about 7-8 percentage points for all other topics groups. These correspond to about 25% of the value taken by the mean for groups 2, 4 and 5, to about 30% for group 1, and to more than 40% for geometry topics (group 3). Similarly, the ratio between the standard deviation and the mean is around 30% for groups 1 and 3 and never above 25% for all other groups.

[Table 2 here]

Panel B reports descriptive statistics on the content composition of the state curriculum standards as measured by the percentage of test items covering each topic group in the end-of-year grade 4 state test. As reported in Column 1, the curriculum standards averages are very similar to those of class time allocation, thus suggesting a potential alignment between the two. However, descriptive statistics in Panel C show that differences between standards and classroom instruction are in fact significant. Specifically, Column 1 shows that, for each topic, the average absolute deviation between these two is always greater than the standard deviation of class time allocation. This is particularly true for the geometry group, where the value of the mean absolute deviation is 40% larger than the standard deviation.

[Table 3 here]

A potential reason behind this misalignment with the standards could entail adjustments of teaching strategies to the composition of the classroom. As displayed in the upper panel of Table 3, classrooms could vary substantially in their composition as it pertains to students' level of math readiness. Columns 3-5 show that one fourth of the classrooms in the sample have a fraction of low-achieving (1^{st} tercile) students below 0.14 while another one fourth display values above 0.46. This range is a little higher for the fraction of high-achieving students, with 25% of the classrooms having

[20] The SEC allows class time to be allocated in 183 topics combined with five possible levels of "cognitive" demand, for a total of 915 cells. The choice of the five topic groups was inspired by the Common Core standards classification.

14.3% or less students performing on the higher part of the distribution and another 25% with 52% or more. Classrooms tend to have more similar percentage of students with initial knowledge falling in the middle of the distribution, with the first and third quartiles being 24.1% and 41.7%, respectively. The distribution of classroom composition may depend on the implementation of ability tracking in the schools. In particular, schools can choose to track students at different intensity levels, where the level of intensity is the importance given to students' prior performances when assigning them to classrooms. One way to measure tracking intensity in a school entails computing the share of total variance of baseline test scores due to between-classroom variation. Indeed, schools that track students would display higher shares of between-classroom variation due to higher classroom homogeneity and, as a result, a lower share of within-classroom variance. At the extreme, the share of between-classroom variance can range from zero or close to zero to a maximum given by the value taken by the highest level of tracking intensity, represented by the scenario in which students are sequentially assigned to classrooms from low to high baseline knowledge.[21] The lower panel of Table 3 reports descriptive statistics of tracking intensity for 4^{th} grade math classrooms across all schools participating to the the MET study in Year 1. The comparison between the average share of between-classroom variation in the data with the tracking policy configuration suggests that schools apply a very low level of tracking. This is confirmed by the descriptive statistics on the ratio between these two shares (third row), which gives a scale of the degree of tracking relative to the most extreme case. With 1 being the maximum, schools display an average degree of tracking of about 0.12, with 75% of all schools scoring at or below 0.145. Figure B.3 in the Appendix gives a broader look to the level of tracking implemented by the schools. As seen, the distribution of tracking intensity observed in the data is shifted to the left of the same distribution under the highest degree of tracking possible within each school.[22] This result is not surprising, as schools are much less likely to track in lower grades. Indeed, national statistics show that about 30% of US schools implements any sort of ability tracking in 4^{th} grade math classes, and only 5% for 4^{th} grade reading classes.

[Table 4 here]

Table 4 provides descriptive statistics of the measures of used in the estimation of the model. Baseline knowledge is measured by standardized score of the 3^{rd} grade math test administered by the state at the end of each school year, rescaled to have mean 500 and standard deviation 100 at the district level.[23] As seen, the mean and standard deviation in the final sample are 510.20 and 95.40, respectively, hereby showing slightly higher average math performances (with lower dispersion) compared to the district average. There are two measures of end-of-year knowledge. The first measure is

[21] The value of tracking intensity attains zero whenever it is possible to allocate students such that average test scores are identical across classrooms within the school, which is possible only for special configurations of the within-school sample.

[22] Further analysis show that the distribution of tracking intensity displayed in the data is very similar to the one obtained randomly assigning students to classrooms.

[23] The rescaling follows the convention of national and international education athorities as well as prior empirical work.

the end-of-year 4^{th} grade state test-score in mathematics administered by the state, also rescaled to have mean 500 and standard deviation 100 at the district level. As shows by Table 4, the mean and standard deviation in the final sample are 505.38 and 96.79, respectively. Hence, similarly to the 3^{rd} grade scores, students in the final sample perform better in math in 4^{th} grade than the district average and display a slightly lower variability. The second measure of end-of-year knowledge is the percentage of correct questions in the Balanced Assessment of Mathematics (BAM) test, which was administered by the MET study staff to the participating students. Table 4 shows that, on average, students answer about 55% of the questions correctly, with a standard deviation of about 21%.

There are 11 measures for latent teacher ability, including three Classroom Assessment Scoring System (CLASS), two Framework for Teaching Mathematics (FFTM), and two Mathematical Quality of Instruction (MQI) scale scores from video-recorded lessons, as well as four measures from the student survey. The CLASS scores are measured on a scale of 1 to 7 and include: *(i)* the behavior management score, which evaluates teacher's ability to set clear behavior expectations, to prevent and redirect students misbehavior, and to obtain students' compliance; *(ii)* the content understanding score, which refers to both the depth of the lesson's content and the teacher's ability to help students in understanding the framework and key ideas of the topic taught; *(iii)* and the productivity score which measures teacher's level of preparation for the lesson as well as her ability to maximize learning time and to set clear routines and instructional expectations. As for the FFTM scores, measured on a scale of 1 to 4, the management of class procedures score measures the degree of smooth functioning of the classroom, whereas the management of student behavior score evaluates the teacher's ability to manage student conduct and to respond to their misbehavior. Differently from CLASS and FFTM measures, the MQI scores assess the pedagogical knowledge and preparation of the teacher necessary to teach mathematics. Specifically, the richness of mathematics score refers to the teacher's ability to explain mathematical ideas as well as to draw connections and illustrate different aspects of math concepts, while the mathematical knowledge for teaching (MKT) score measures the overall teacher's knowledge in the specific area of mathematics taught in 4^{th} grade. Finally, the last four measures are class-level averages of student evaluation scores on the teacher's ability to deliver instruction, where each single student response is an ordinal variable converted to a measure taking values from 1 to 4. As seen, teachers have pretty good evaluations in terms of their ability to explain concepts (with averages higher than 3), whereas they tend to get lower ratings with respect to their ability to control the class behavior. Similarly to the last four measures of teaching ability, the 8 measures of latent teacher effort come from student evaluations and take values on a scale of 0 to 4. Specifically, these measures refer to the teacher's level of feedback and motivational support provided to the students, as well as to her level of effort in trying to avoid that students fall behind during the lesson. As shown by the third panel of Table 4, teachers average score is higher than 3 in five out of eight measures, while lower scores are reported in terms of teachers' effort to not waste time in class, to summarize the lesson, and to write feedback on homework and exams. As for teacher preference for the adherence of instruction to the curriculum standards, the measures include the teacher's self-reported degree at which her school administrators require teachers to adhere to the standards and the frequency at which she uses curriculum standards documents (both 5 categories). As shown by the mean values above 2, the majority of teachers report a high degree of required adherence to the standards by administrators as well as a frequent use of standards documents.

At the student level, measures of student effort are represented by the answers to 4 questions included in the student survey. Students report a quite high level of effort in school activities, with more than 40% of the sample declaring to always do their best work in class, to never give up when work gets hard, to never take it easy not trying to do their best, or to complete all the homework assigned. Measures of student preference for own knowledge include response to questions on whether the student finds school work interesting and/or enjoyable, and on whether the student reads at home daily. The last panel of Table 4 displays a relatively higher heterogeneity in the responses compared to student effort. More than half of the students find school work interesting either all or most of the time (25% and 31%), while 29% of them finds it interesting sometimes and 15% uninteresting. About 45% of the students think that school work is never or mostly never not enjoyable (29% and 16%), while 29% report to not enjoy school work either always or at least most of the time (12% and 17%). Finally, a 65% of students report to read at home almost every day, 22% do that sometimes, and only 13% rarely or never reads at home.

[Table B.1 here]

A necessary condition for the identification of the model's parameters entails the non-independence between measures of each latent variable. To this end, Table B.1 in the Appendix reports the correlation matrices of the measures described in Table 4. Specifically, the correlation between continuous variables is measured by the Pearson's correlation coefficient, while the Pearson's chi-squared statistics is computed to assess the association between categorical variables. Almost all pairs of measures display statistically significant correlations, with the exception of the student-survey measures of teacher ability which more than half of the times are unrelated to the CLASS, FFTM, and MQI scale scores.

5. Estimation Results

5.1 Parameter Estimates

Table 5 reports the estimated parameters of the production and utility functions (1) and (3). Estimates of the class time taste shocks covariance matrix and of the measurements and exogenous latent factors equations are instead reported in Table B.4 in the Appendix. Column (1) shows that a positive value of the parameter δ_0, which captures both knowledge depreciation between grades 3 and 4 and a normalization, since K_{0ti} and K_{1ti} are different cardinal measures of math knowledge. The coefficients converting input units, $(\delta_{1d})_{d=1}^{5}$ are all positive and statistically different from 0 at the 0.01 level, with an average value of about 0.0005.[24] The estimates show that baseline knowledge and teacher ability are both positively related to end-of-year knowledge, with values of 1.1288 and 0.3414, respectively. The total elasticity of teacher effort, $\gamma_{20} + \gamma_{21} N_t$, is estimated to be about 0.0139 for an average class size N_t of 23 students. Both γ_{20} and γ_{21} are statistically significant a the 0.01 level, with the negative sign of γ_{21} suggesting that teacher effort is less productive in larger classes, with a decrease in elasticity of about 0.0004 points for each additional student. Finally, the elasticity of student effort is relatively small and imprecisely estimated. Column (3) reports the estimates capturing direct peer-to-peer spillovers together with the standard deviation of the random shock.[25] Given the normalization $\pi_{q2} = 0$ for $q = 1, 2, 3$, each estimate is interpreted as the effect of an increase in the fraction of classmates in tercile $k = 1, 3$ in response to an identical decrease in the fraction of students in the 2^{nd} tercile. As seen, an increase in the share of classmates in tercile 1 (relative to a decrease second tercile students) tends to harm students in the first two terciles, although both π_{11} and π_{21} are both imprecisely measured. Interestingly, students in the highest tercile seem to benefit more from low-achieving peers than those in the middle range of the distribution, as shown by $\pi_{31} > 0$. On the other hand, the effect of an increase in the share of high-achieving classmates (in response to an equal decrease in the share of 2^{nd} tercile peers) on the achievement of students in the first tercile has negative sign though not statistically significant at canonical levels. Both students in the second and third tercile benefit from a higher share of third-tercile students in the classroom, with high-achieving students experiencing the highest spillovers from peers with similar prior knowledge.

Panel B in Table 5 reports the elasticities of class time inputs conditional on initial knowledge tercile. For all terciles, the most productive topics are those related to either unit conversion and measurement or fractions and decimals, with estimated values all above 0.25. Although likely the most difficult topic for fourth graders, time spent teaching fractions and decimals seem very productive for students in the 1^{st} tercile ($\eta_{41} = 0.385$), especially compared to students in the 2^{nd} tercile. Time allocated to unit conversion and measurement has a higher elasticity on students in the middle range of baseline knowledge ($\eta_{52} = 0.356$), while it has less of an effect on students in the top tercile ($\eta_{53} = 0.257$). The topics displaying the lowest elasticities are those related to geometry, with estimated values never above 0.07. Students in the top terciles seem to benefit the most from geometry topics, while the estimates for students with baseline knowledge in the 1^{st} and 2^{nd} terciles are imprecisely measured. The more basic topics of place value, rounding, addition, and subtraction, also display low elasticities across all terciles. Interestingly, time allocated to this topic group is

[24] District-specific values are not reported due to confidentiality agreements.

[25] Both K_{0ti} and K_{1ti} have been divided by 100 for the estimation. Hence, the parameters have to be interpreted accordingly.

more productive for higher students in higher terciles. Finally, the elasticities of time spent teaching multiplication and division are very similar across terciles, with estimates ranging between 0.247 and 0.270.

[Table 5 here]

The utility function parameters are reported in Panel C of Table 5. As shown in Column (1), teachers attach the highest value to achievement gains of students with low baseline knowledge (ω_1^1) and the lowest to students with high initial knowledge (ω_1^3). Hence, teachers exhibit a higher preference for compensatory teaching aimed at fostering the learning gains of students starting with lower levels of math knowledge. In particular, a unit increase in math achievement from of a student with low initial knowledge rewards the teacher about 2.2 times the same increase for a student whose level of prior math knowledge falls in the third tercile.[26]. Moreover, the estimates of (ω_1^2, ω_2^2, and ω_3^2 suggest that teachers tend value more the achievement of both black and Hispanic students, while they attach a slightly lower value to female students. Column (3) shows the estimated parameters on teacher preferences over class time allocation. Each $\tilde{\alpha}_{1j} \equiv \alpha_{1j} - \alpha_{15}$ (for $j = 1, \ldots, 4$) captures the utility a teacher gets from reallocating 1% of class time from unit conversion and measurement to topics in group j, while holding fixed the value of all the other terms in the utility function. The sign of the estimates suggest that teachers have a general preference for time spent away from unit conversion and measurement topics, with the only exception of fraction and decimals (topic 4). In particular, teacher seem to get much more utility from time spent teaching geometry (topic 3) compared to all other topics. On the other hand, the parameters on topics 2 and 4 are not statistically different from zero. Moreover, a likelihood ratio test fails to reject the null that $\tilde{\alpha}_{11} = \cdots = \tilde{\alpha}_{14} = 0$.[27] Finally, the relatively high and statistically significant estimate of α_{203} suggests that teachers tend to adhere to the standards associated with geometry topics. Otherwise, the low estimates of the parameters $(\alpha_{20j})_{j=1}^{5}$ and α_{21} indicate that teachers do not bear significant costs if they teach away from what suggested by the education authorities in all other topics.

5.2 Within-Sample Model Fit

Columns (1)-(4) in Table 6 compare the means and standard deviations obtained from simulations of the estimated model with the actual values observed in the data. Overall, the model fits the data well. The predict mean and standard deviation of both baseline and end-of-year knowledge (measured, respectively, by the 3^{rd}-grade math state test and the BAM test score) very close to the actual ones. Indeed, the model slightly overestimates statistics on baseline knowledge, with differences

[26] A LR test rejects the null hypothesis that these weights are all equal.

[27] Notice that $\tilde{\alpha}_{11} = \cdots = \tilde{\alpha}_{14} = 0$ is equivalent to $\alpha_{11} = \cdots = \alpha_{15}$, where each α_{1j} could well be different zero. Hence, the test fails to reject the null that teachers value time spent on each topic the same.

of about 1% for the mean and to 2% for the standard deviation, respectively, while it moderately underestimates the mean and standard deviation of the BAM score by about 5% and 3%, respectively. As for the class time allocations across topics, the average simulated values are very similar to the respective data means, with the largest difference being in the time spent on place value, rounding, addition and subtraction, whose mean value is 1.0 percentage points higher than in the data. On the other hand, the model tends to systematically over-estimate the standard deviations of class time allocation. In particular, the model is very imprecise in predicting the standard deviation of place value, rounding, addition and subtraction topics, whose prediction is 4 percentage points higher than the one found in the data. As for the other topic groups, the model overestimates the standard deviation by no more than 1.7 percentage points (i.e., fraction and decimals). Yet, although these difference are quite large relative to the standard deviations observed in the data, they do not seem substantial when compared to the mean values. Finally, the model fits very well the sample statistics of all other measures, with minimal discrepancies in both mean and standard deviations.

[Table 6 here]

Column (5) in Table 6 reports the share of total variance due to the latent factor (1 minus the fraction reflecting measurement error). These shares are not reported for baseline knowledge and class time allocation inputs, as these inputs are assumed to be measured without errors. The values reported shows a substantial level of measurement error in the measures used in the analysis. About 53% of the variance of the BAM test score is due to true variation in end-of-year knowledge, while the remaining 47% represents measurement error. The importance of measurement error varies significantly across measures of teacher effort, going from a low of 56% for "Teacher ask questions..." to a maximum of 92.3% and 99.7% for average responses to the questions "The teacher writes feedback on our papers" and "The teacher pushes students to work hard", respectively. All other teacher effort measures entail a degree of measurement error reflecting between 60% and 80% the total variation of the measure. On the other hand, the importance of measurement error is quite uniform across measures of student effort, despite being always as high as 86%. An even wider range of measurement error importance is displayed by teacher ability measures. Indeed, these include very precise measures like those from the FFTM protocol (with only 19% and 21%) as well as a set of very noisy measures represented those collected through the MQI protocol, with shares of measurement error between 87% and 99%. As for the other ability measures, those collected through the CLASS protocol tend to be relatively accurate with an average of about 40% of their variance reflecting the actual variation in the latent factor. Teacher ability measures based on student survey evaluations, instead, tend to display higher levels of measurement error, ranging between 67% to 80%. Probably the most error-ridden measures presented in Table 6 are those on student preference for own knowledge, where both responses to "School work is interesting" and "School work is not enjoyable" displaying a share of “true” latent factor variance below 8.5%. Yet, a slightly more precise measure is "I read at home almost every day", where measurement error accounts for “only” 70% of its total variance. Finally, the measures on teacher preference for adherence to the standards show very different levels of measurement error. Indeed, while variation in responses to the question "Administrators require rigid adherence to the standards" are due to measurement error for almost 80% of their value, the measure "I frequently refer to and use information found in standards documents" mostly reflect true variation in the latent factor.

5.3 Out-of-Sample Validation

I exploit the second-year data of the MET study to perform an out-of-sample validation of the model. In particular, this exercise entails using the estimated parameters to predict second-year outcomes given both teacher and student initial conditions. This sample includes all the teachers participating to the second year of the study who were randomly assigned to a classroom within the same school.[28] The randomization was performed by MET researchers in order to correct for the potential bias in the estimates of teacher value-added caused by the non-random assignment of teachers to classrooms, especially when the latter is based on unobservable student or teacher characteristics. In the theoretical framework of this paper, the non-random assignment of teachers in Year 1 of the study would bias the effect of the teacher inputs A_t, $\boldsymbol{\tau}_t$, and e_t on end-of-year knowledge if these were correlated with omitted inputs even after controlling for prior achievement. Therefore, this exercise allows me to check indirectly whether the model specification is able to capture the variation underlying the teacher assignment mechanism.

Table B.6 compares descriptive statistics of selected variables from the sample used to estimate the model and the one used to perform the validation exercise. There are several major differences between students and teachers in these two samples. First, students in the second-year sample display lower initial and end-of-year knowledge in math, with test scores being about 0.13σ lower than the first-year sample. Second, students are younger in the Year 2 sample, with an average age lower by about 0.6 (equivalent to about seven months). Finally, teachers have, on average, one less year of experience teaching in the district, classrooms are larger in size, and their composition tend to be more skewed towards students with low levels of baseline knowledge. Indeed, while the fraction of 2^{nd} tercile students is very similar compared to Year 1, classrooms have on average 5 percentage points more students in the 1^{st} tercile and about 5 percentage points less students in the 3^{rd} tercile. All other student and teacher characteristics are very similar between the two samples.

[Table 7 here]

Table 7 compares predicted and actual means and standard deviations using the second year sam-

[28] Not all teachers participating to the study were randomly assigned to a classroom in Year 2. See the MET documentation for further details.

ple. As shown by the values reported, the model does a very good job in predicting all the outcomes outside of the sample used in estimation, and virtually the whole analysis on the discrepancies between data and predictions made in Section 5.2 holds in this case as well. The only exception is represented by the predicted means and standard deviations of the two measures on teacher preference for adherence to the standards, whose values severely underestimate the actual values by more than 50%. Finally, despite the absence of data on class time allocation in the second-year sample, which does not allow me to assess the model fit, it can be noticed that the model predicts the class time allocations to be different in the second year of the study. In particular, according to these simulations, teachers spend more time teaching place value, rounding, addition, and subtraction, multiplication and division, and fractions and decimals.

5.4 Discussion: Teacher Rewards and Educational Incentives

As discussed in the section above, the estimated weights $\left(\omega_1^q\right)_{q=1}^3$ in Table 5 suggest that teachers in the school districts represented in the sample tend to value more the learning gains of students at the bottom of the distribution. This result is highly consistent with the incentives that teachers face from the US educational system. At the national level, the No Child Left Behind (NCLB) act was in full regime in 2009 and 2010, when the MET data were collected. In particular, NCLB tied the disbursement of funds directed to Title I schools and other local education agencies (LEA) to their student performances as measured by the so-called Adequate Yearly Progress (AYP).[29] The latter represents the improvement the school needs to attain in the share of students performing above proficiency level in order to reach the ambitious goal of 100% proficiency by the end of 2014. Moreover, additional steps are taken to improve eligible schools that fail to attain the AYP for multiple years, such as staff replacements, providing students with a transfer option, and even closing the school for good. These rules are clearly aimed at providing schools with the incentive to focus their efforts on pupils below the proficiency level. In particular, recent empirical evidence suggests that educators concentrated their effort on students a the margin of the proficiency cutoff (see Macartney et al., 2021). The schools represented in the MET data are particularly exposed to the NCLB incentives, with the districts having between 70%-to-90% Title I schools.

Furthermore, almost all school districts represented in the sample run their own local teacher performance-pay programs during the MET study.[30] In the NYC Public Schools district, about 200 low-performing schools were randomly selected to participate in the School-Wide Performance Bonus Program (SPBP).[31] Each of these schools can earn up to $3000 per staff member, represented

[29] Schools are eligible to receive Title I funds if they have a representation of low income students (as measured by the share of students receiving reduced-price or free lunch) above a specific threshold.

[30] The only district in the sample not running any such program is Memphis City Schools.

[31] This program was part of a randomized controlled trial in which 200 schools were randomly assigned to the program out of a total pool of 400 low-performing schools . The total number of schools in the NYC district in 2009 was about 1,600. (For more details about the program and its effects see Springer, 2011; Fryer, 2013).

by the United Federation of Teachers (UFT), if it attains specific performance targets. Targets are expressed as a score reported yearly by a Progress Report Card. Specifically, 55% of this score comes from student performances, which in elementary and middle schools are measured by the average change in state proficiency ratings (based on the NY state exam) and by the percentage of students making a year of progress among the bottom third. Hence, somewhat similarly to NCLB, schools participating in SPBP have an incentive in focusing their attention on students at the lower quantiles of the distribution. A similar incentive seems to be envisioned by the Merit Awards Program (MAP) run by the state of Florida starting from 2007, which rewarded teachers in the top quartile of an assessment which depends for a 60% on either student proficiency, student learning gains, or both. Hillsborough County Public Schools participated in MAP in during the MET study, with rewards ranging between 5% to 10% of a teacher salary. Different mechanisms were instead envisioned by the performance-pay programs in Charlotte-Mecklenburg (the TIF-LEAP program) and Denver Public Schools (ProComp program). Indeed, in both districts teacher performances were partly measured with respect to specific objectives on student learning outcomes that the teacher establishes at the beginning of the year with a school leader or with some education professionals.[32] These goals could include improvement in the achievement of specific segments of the classroom, like low-performing or disadvantaged students. Finally, the Dallas Independent Schools District (ISD) Performance-Pay Program implemented in 2008 awarded teachers whose estimated value-added measure (referred by the program as Classroom Effectiveness Index) above the 70^{th} percentile in the same district. However, differently from the programs discussed above, it is not clear whether and how incentives based on value added measures would incentivize educators to concentrate their effort on specific groups of students, although there are plausible mechanisms that could generate such a behavioral response.[33]

6. Counterfactual Analysis

6.1 Ability Tracking and Teacher Assignment Mechanisms

The parameter estimates discussed in Section 3 allow me to investigate the distributional impact of tracking on student achievement. In order to do that, I simulate several counterfactual scenarios where I reassign students to classrooms based on their prior knowledge, K_{0ti}. This policy experiment is feasible since, as already described in Section 4, there is no evidence of tracking in 4^{th} grade math classes among schools participating to the MET study. In particular, I choose to simulate the most

[32] These objectives are called Student Learning Objectives (SLO) in Charlotte-Mecklenburg and Student Growth Objectives (SGO) in Denver Public Schools.

[33] For instance, a possible mechanism could entail the fact that students at the bottom of the distribution have much more room for improvement, as shown for instance by the unusually high impact of a individualized instruction intervention in Gambia assessed by Eble et al. (2021). On the other hand, one can argue that low-achieving students are generally less motivated and, hence, less responsive to teacher and school efforts. In this scenario, focusing on stronger and more motivated students could be more productive.

extreme configuration of ability tracking, i.e., by ranking students from lowest to highest K_{0ti} within each school and then assigning them sequentially to each classroom. To have a full representation of all the 4^{th} grade teachers in each school, I perform this policy experiment using the full sample of 4^{th} grade math classes in the first year of the MET study.[34] In the simulation, I keep the same number of classrooms as the original data and I change class sizes to be the same within each school.

Once students are assigned to classrooms, a second-order issue involves the assignment of teachers to each track. As I will show below, the choice of the teacher assignment mechanism turns out to be crucial for the distributional impact of tracking on achievement. In practice, there is evidence that schools make these assignments in a non-random fashion, usually based on both teacher and student characteristics. For instance, Kalogrides et al. (2013) show that, in a large school district in Florida, schools tend to assign more educated and/or experienced teachers to high-achieving students, and either female or minority teachers to lower-achieving ones. This is usually due to the accumulation of both organizational and social capital by more experienced teachers, which makes them more influential in terms of the assignment decisions (Grissom et al., 2015). To the extent that teacher's experience is correlated with instructional ability, these assignment patterns might either overlap or contrast with the specific goals the school wants to pursue in terms of student achievement. In the present analysis, I take a closer look at this issue and use teacher's ability as a discriminant for their assignment to different tracks within each school. I then compare the resulting outcomes to allocation mechanisms based on years of experience. Specifically, I first simulate ability tracking under three alternative teacher assignment mechanisms: 1) random assignment (RA); 2) higher ability teachers to higher tracks and low-ability teachers to lower tracks (*positive assortative matching*, or PAM); 3) higher ability teachers to lower tracks and lower ability teachers to higher tracks (*negative assortative matching*, or NAM).

Plot (a) in Figure 1 shows the impact of tracking on end-of-year knowledge for students at each tercile of the prior knowledge distribution and across different teacher assignment mechanisms.[35] As seen, the overall effect of tracking on achievement ("All terciles" group) is positive and similar across teacher assignment mechanisms, with values ranging between 0.015SD (NAM) and 0.02SD (PAM). Yet, the effects are very heterogeneous across both terciles and assignment mechanisms. When assigning teachers at random (green bars on the right of each tercile group), we see that tracking yields a nearly zero negative effect on 1^{st} tercile students and a decrease in achievement of 2^{nd} tercile students by about 0.036SD, while students in the top tercile experience a quite significant increase in achievement of about 0.10SD. On the other hand, assigning teachers with higher ability to upper

[34] I therefore include teachers not entering the sample used to estimate the model (i.e., those who did not take the SEC survey, and therefore did not report information on their class time allocation).

[35] Results are based on a total of 8,000 simulated classrooms. Given the missing information on class time allocation for a large part of the teachers in the sample, the baseline values are also simulated given the assignments of students and teachers to classrooms in the data.

tracks with PAM (and, therefore, allocating the low ability teachers to students at the bottom of the distribution) yields results with same sign but larger magnitudes (light blue bars in the middle of the tercile groups). Indeed, for both 1^{st} and 2^{nd}-tercile students tracking is even more detrimental when assigned to lower quality teachers, with their achievement decreasing by about 0.037SD with respect to the original classroom assignments. Conversely, students at the top of the distribution experience an even higher improvement in achievement when assigned to high-ability teachers, with an increase in end-of-year knowledge of about 0.16SD. Finally, the distributional effect of tracking is significantly different when the best teachers are, instead, assigned to students in the lowest part of the distribution (i.e., negative assortative matching). Indeed, Figure 1 shows that students at both the top and the bottom terciles of the distribution benefit from tracking, with an increase in end-of-year knowledge of 0.05SD and 0.027SD, respectively. Students in the middle of the distribution, instead, experience almost the same decline in knowledge as under RA and PAM. As a result, depending on how teachers with different instructional ability are assigned to classrooms, tracking can either hurt, benefit, or leave unaltered the performance of students at the bottom of the distribution. On the other hand, high-achievement students benefit from tracking no matter the teacher assignment mechanism, although the magnitude of these positive effects can vary substantially depending on the ability of the teacher they are assigned to.

Finally, Panel (b) in Figure 1 reports the effect of tracking on achievement when, rather than on ability, teachers are assigned to tracks based on their years of teaching experience. The figure illustrates a pattern somewhat similar to the ability-based assignments, although with different magnitudes. In particular, while the effect of tracking still differs substantially across terciles, it also changes very little under different teacher assignment mechanisms. A potential reason behind this result is that years of experience are, generally, weakly correlated with teacher ability. Indeed, as shown in Table B.2, the estimates of the parameters in $\boldsymbol{\beta}_1^A$ on teacher experience are not statistically significant. These findings are consistent with a common pattern found in the literature, according to which teacher quality is usually poorly explained by characteristics like experience or educational attainment (see e.g., Rivkin et al., 2005).

6.2 Teachers' Instructional Adjustments and Peer Spillovers

On top of the specific teacher assignment mechanism employed, the heterogeneity in the effect of tracking across terciles is driven by changes in the composition of the classrooms through teachers' instructional adjustments as well as direct peer-to-peer spillovers. Teachers can adjust instruction by altering both the allocation of class time across topics and the amount of effort they exert. Table 8 reports the average allocation of class time across topics delivered to students at different terciles under both the baseline (non-tracking) and the tracking scenarios. In particular, the present analysis focuses on tracking in the case of random assignment of teachers to classrooms. Results under either type of assortative matching are similar and, therefore, are omitted. Columns (1) and (2) show the average values of $\boldsymbol{\tau}_t$ delivered to students at different terciles under the baseline and tracking scenarios, respectively. Column (3), instead, reports the allocation of class time tailored to student's prior knowledge at each specific tercile, computed as $\widetilde{\boldsymbol{\tau}}^q = \overline{\tau} \times \boldsymbol{\eta}_q$ for $q = 1,2,3$. The comparison of these three columns shows that, under tracking, students in the first tercile tend to receive values of $\boldsymbol{\tau}_t$ slightly closer to $\widetilde{\boldsymbol{\tau}}^q$, while for students at higher terciles there is no clear pattern on how class time allocation adjusts. These results are likely a by-product of the higher value teachers attach to the achievement of students in the first tercile. Hence, teachers have a higher incentive to tailor instruction to students' needs when assigned to lower tracks. On the other hand, other factors like curriculum standards and preferences for time spent on specific topics seem to drive the adjustments of $\boldsymbol{\tau}_t$ for teachers assigned to higher tracks.[36]

Panel (a) of Figure 2 shows the effect of tracking on the level of instructional effort experienced by students at different terciles for different teacher assignment mechanisms. Under tracking with negative assortative matching (blue bars at the left of the tercile groups), students at the bottom of the distribution see an increase of effort by the teachers they are assigned to equivalent to about 0.7SD, while students at the top tercile experience a drop in instructional effort of the same magnitude. Similar results are found when teachers are randomly assigned to classrooms, although with much smaller magnitudes (grey bars on the right of each tercile group), with the changes in teacher effort for students at the bottom and top terciles being +0.2SD and -0.25SD, respectively. Conversely, under positive assortative matching students in the bottom tercile see a decrease in the instructional effort of their teacher of almost 0.2SD, while third-tercile students experience an increase of about 0.1SD. As for students in the second tercile, teacher effort increases by less than 0.1SD no matter how teachers are assigned. These result are a clear reflection of interactions between teachers' ability and the rewards they get for the achievement of different students. On one hand, teachers tend to increase (decrease) effort when assigned to students whose achievement they value the most (least). On the other hand, more able teachers tend to exert more effort because their instruction is generally more productive. As a result, these effects can either magnify or offset each other depending on how teachers with high or low ability are assigned to lower or higher tracks.

A final channel through which changes in classroom composition brought about by tracking influence student outcomes is direct peer-to-peer spillovers. Table 9 displays how the total effect of tracking on the achievement of students at different terciles is the sum of the effect of both direct and indirect peer-to-peer spillovers. The sign and magnitude of direct peer spillovers reflect those of the estimated parameters π_{qk}, and their values are, by construction, constant across teacher assignment mechanisms. In particular, tracking generates large positive direct spillovers for students in the

[36] Although similar, these effects can be more or less pronounced under depending on the teacher assignment mechanism. In particular, given the complementarity between $\boldsymbol{\tau}_t$ and teacher ability, negative (positive) assortative matching leads teachers assigned to lower tracks to teach closer (further) to the level of instruction tailored to students' knowledge.

top tercile, and negative ones for all other students. On the other hand, Table 9 shows how indirect spillovers originating from teachers' instructional adjustments are able to either reinforce or offset the effect of direct peer effects, where these adjustments are represented by changes in class time allocation and instructional effort discussed above. Specifically, assigning low ability teachers to lower tracks and high ability teachers to higher tracks (PAM) further widen the achievement inequality already generated by direct peer spillovers. On the contrary, assigning better teachers to students in lower tracks (NAM) yields high enough positive indirect effects to both compensate and reverse the sign of the overall effect of tracking on students in the bottom tercile, while students in the top tercile experience lower, but still positive, total effects.

6.3 The Impact of Curriculum Standards on Instruction and Achievement

Since the early 1990s, many education policies have been involved in the establishment of educational standards. The main objective of these standards is to establish a common benchmark for student proficiency across schools, as well as to provide teachers with guidelines on how to structure their curricula and pace of instruction. Despite their importance in the education policy agenda, empirical evidence on the effectiveness of curriculum standards in raising student achievement is still inconclusive. Hence, a first-order question is whether the existing curriculum standards are set at a level which would actually foster student achievement. To do that, I simulate a counterfactual where I impose teachers to teach according to the standards in their state. Formally, I set $\boldsymbol{\tau}_t = \varphi_t$ for each s and t. The simulation results are reported in Table 10. Comparing the counterfactual scenario with the status quo (where teachers are free to choose their time allocation) it is possible to see that teaching according to the state-level standards would be slightly detrimental for students along the entire distribution of prior achievement. Hence, these results suggest that curriculum standards in the 5 states represented in the MET data are, in general, not well-suited to the students' level of prior knowledge.

7. Conclusion

This paper explores the relationship between instruction, classroom composition, and student knowledge accumulation by developing and estimating an equilibrium model of endogenous instruction and student effort. Teachers maximize their utility by choosing how much effort to exert in class as well as the allocation of instructional time among different topics. The model allows teachers to attach different weights to the achievement of students with different levels of prior knowledge, race, or gender. Students also maximize their utility by choosing learning effort. The equilibrium is modeled as the outcome of a static game of complete information. The model also specifies a technology of knowledge production which allows class time allocation to have a differential impact on the achievement of students at different levels of the distribution of prior knowledge, and to incorporate direct peer-to-peer spillovers not mediated by the teacher's behavior. For the empirical part of the analysis, I use a sub-sample of fourth grade math classes from the first year of the MET project data. The estimation is carried out through maximum simulated likelihood. Estimates of the model suggest that teachers attach higher values to the achievement of students with lower levels of initial knowledge. These results are consistent with the incentives provided by the US education system at both the federal and local level in the past two decades, especially from policies like NCLB. Moreover, students with different baseline knowledge display different learning profiles, as shown by the difference in the quantile-specific class time elasticity parameters. The model fits the data well both within and out-of-sample. In particular, the model predicts accurately both student-level and teacher-level outcomes from the second year of MET study, where teachers were randomly assigned to classrooms within each school. These results suggest that the estimates are not significantly affected by the potentially non-random assignment of teachers in the first year of the study.

The counterfactual analysis involves the implementation of ability tracking within each school. I find that the distributional effects of tracking are heterogeneous and depend heavily on how teachers with different ability are assigned to classrooms. In particular, while students at the top of the distribution are always positively affected by tracking (with a peak increase in achievement when assigned to high-ability teachers), assigning high-ability teachers to lower tracks yields positive effects on the achievement of students in the first tercile, thus offsetting the negative effect stemming from the reduction in peer quality. Further analysis shows that teachers respond to tracking by both better tailoring instruction to the students' readiness level and exerting more effort when assigned to lower tracks. Indeed, disentangling the effect of tracking into its direct and indirect peer-to-peer spillovers components, I find that, while students at the bottom of the distribution are those more affected by the negative direct spillovers from lower-quality peers, they are also those benefiting the most from teachers' instructional adjustment after the implementation of tracking. This study contributes to a long-standing discussion on the distributional effects of tracking on students with different levels of prior achievement. In particular, my results highlight the trade-offs generated by tracking when accounting for the endogenous response of teachers, their assignment to different tracks, as well as direct peer-to-peer spillovers. A main takeaway of these results is that, with the right combination of incentives (i.e., teacher rewards) and resources (e.g., high-quality teachers), tracking can benefit disadvantaged students despite the lower peer quality.

The present framework can be expanded in several directions. A natural first step could be to explore what drives the observed heterogeneity in the rewards that teachers attach to the achievement of different students. This has potential implications from a policy standpoint, as it would allow educational authorities to understand the extent to which they are able to incentivize teachers in order to achieve specific policy goals. Moreover, there is a variety of other factors characterizing teacher instructional decisions that are not captured by the mere allocation of time among topics and by the specific measures of instructional effort used in this study. Indeed, broadening the teacher's choice set by including other dimensions of instruction could help uncover new facets of teacher's behavior, especially as related to their interaction with heterogeneous students (see e.g., Aucejo et al., 2021). Further extensions of the present framework could also entail the inclusion of social interactions in the spirit of Blume et al. (2015) or Conley et al. (2018). In particular, including direct peer-to-peer spillovers as equilibrium outcomes would improve on the common specification entailing mechanical peer effects embedded in the education production function. Finally, upon availability of more comprehensive data, the model can be extended to incorporate endogenous parental response to schools, teachers, and peers, whose importance has been highlighted by a growing body of empirical research (see e.g., Fu and Mehta, 2018; Agostinelli, 2018). All these extensions are left for future research.

References

Agodini, Roberto and Barbara Harris, "How four elementary math curricula perform among different types of teachers and classrooms," 2014. Working Paper.

Agostinelli, Francesco, "Investing in children's skills: An equilibrium analysis of social interactions and parental investments," 2018. Working Paper.

— **and Matthew Wiswall**, "Identification of dynamic latent factor models: The implications of renormalization in a model of child development," 2016. NBER Working Paper 22441.

Andrew, Michael D, Casey D Cobb, and Peter J Giampietro, "Verbal ability and teacher effectiveness," *Journal of Teacher Education*, 2005, *56* (4), 343–354.

Argys, Laura M, Daniel I Rees, and Dominic J Brewer, "Detracking America's schools: Equity at zero cost?," *Journal of Policy Analysis and Management*, 1996, *15* (4), 623–645.

Aucejo, Esteban M, Patrick Coate, Jane Cooley Fruehwirth, Sean Kelly, Zachary Mozenter, and Bates White, "Teacher effectiveness and classroom composition: Understanding match effects in the classroom," 2021. Working Paper.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying education: Evidence from two randomized experiments in India," *The Quarterly Journal of Economics*, 2007, *122* (3), 1235–1264.

Barlevy, Gadi and Derek Neal, "Pay for percentile," *American Economic Review*, 2012, *102* (5), 1805–31.

Ben-Porath, Yoram, "The production of human capital and the life cycle of earnings," *Journal of Political Economy*, 1967, *75* (4, Part 1), 352–365.

Betts, Julian R and Jamie L Shkolnik, "The effects of ability grouping on student achievement and resource allocation in secondary schools," *Economics of Education Review*, 2000, *19* (1), 1–15.

Blume, Lawrence E, William A Brock, Steven N Durlauf, and Rajshri Jayaraman, "Linear social interactions models," *Journal of Political Economy*, 2015, *123* (2), 444–496.

Bodovski, Katerina and George Farkas, "Mathematics growth in early elementary school: The roles of beginning knowledge, student engagement, and instruction," *The Elementary School Journal*, 2007, *108* (2), 115–130.

Booij, Adam S, Edwin Leuven, and Hessel Oosterbeek, "Ability peer effects in university: Evidence from a randomized experiment," *The Review of Economic Studies*, 2017, *84* (2), 547–578.

Brock, William A and Steven N Durlauf, "Interactions-based models," in "Handbook of Econometrics," Vol. 5, Elsevier, 2001, pp. 3297–3380.

Burgess, Simon M, "Human capital and education: The state of the art in the economics of education," Technical Report 2016.

Carlana, Michela, "Implicit stereotypes: Evidence from teachers' gender bias," *The Quarterly Journal of Economics*, 2019, *134* (3), 1163–1224.

Conley, Timothy, Nirav Mehta, Ralph Stinebrickner, and Todd Stinebrickner, "Social interactions, mechanisms, and equilibrium: Evidence from a model of study time and academic achievement," Technical Report, National Bureau of Economic Research 2018.

Connor, Carol McDonald and Frederick J Morrison, "Individualizing student instruction in reading: Implications for policy and practice," *Policy insights from the behavioral and brain sciences*, 2016, *3* (1), 54–61.

_ , Michèle MM Mazzocco, Terri Kurz, Elizabeth C Crowe, Elizabeth L Tighe, Taffeta S Wood, and Frederick J Morrison, "Using assessment to individualize early mathematics instruction," *Journal of School Psychology*, 2018, *66*, 97–113.

Cuesta, José Ignacio, Felipe González, and Cristian Larroulet Philippi, "Distorted quality signals in school markets," *Journal of Development Economics*, 2020, *147*, 102532.

Cunha, Flavio, James J Heckman, and Susanne M Schennach, "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 2010, *78* (3), 883–931.

Darling-Hammond, Linda and Peter Youngs, "Defining "highly qualified teachers": What does "scientifically-based research" actually tell us?," *Educational Researcher*, 2002, *31* (9), 13–25.

Del Boca, Daniela, Christopher Flinn, and Matthew Wiswall, "Household choices and child development," *Review of Economic Studies*, 2014, *81* (1), 137–185.

Deming, David J and David Figlio, "Accountability in US education: Applying lessons from K-12 experience to higher education," *Journal of Economic Perspectives*, 2016, *30* (3), 33–56.

Donaldson, Morgaen L, Kimberly LeChasseur, and Anysia Mayer, "Tracking instructional quality across secondary mathematics and English Language Arts classes," *Journal of Educational Change*, 2017, *18* (2), 183–207.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 2011, *101* (5), 1739–74.

Eble, Alex, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Tony Brady, Piotr Gawron et al., "How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in The Gambia," *Journal of Development Economics*, 2021, *148*, 102539.

Engel, Mimi, Amy Claessens, and Maida A Finch, "Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten," *Educational Evaluation and Policy Analysis*, 2013, *35* (2), 157–178.

__ , __ , Tyler Watts, and George Farkas, "Mathematics content coverage and student learning in kindergarten," *Educational researcher*, 2016, *45* (5), 293–300.

Figlio, David N and Marianne E Page, "School choice and the distributional effects of ability tracking: does separation increase inequality?," *Journal of Urban Economics*, 2002, *51* (3), 497–514.

Fryer, Roland G, "Teacher incentives and student achievement: Evidence from New York City public schools," *Journal of Labor Economics*, 2013, *31* (2), 373–407.

Fu, Chao and Nirav Mehta, "Ability tracking, school and parental effort, and student achievement: A structural model and estimation," *Journal of Labor Economics*, 2018, *36* (4), 923–979.

Gershenson, Seth, Cassandra Hart, Joshua Hyman, Constance Lindsay, and Nicholas W Papageorge, "The long-run impacts of same-race teachers," Technical Report, National Bureau of Economic Research 2018.

Gilligan, Daniel O., Naureen Karachiwalla, Ibrahim Kasirye, Adrienne M. Lucas, and Derek Neal, "Educator Incentives and Educational Triage in Rural Primary Schools," *Journal of Human Resources*, forthcoming.

Goldberger, Arthur S, "Structural equation methods in the social sciences," *Econometrica*, 1972, *40* (6), 979–1001.

Grissom, Jason A, Demetra Kalogrides, and Susanna Loeb, "The micropolitics of educational inequality: The case of teacher–student assignments," *Peabody Journal of Education*, 2015, *90* (5), 601–614.

Gromada, Anna and Claire Shewbridge, "Student learning time: A literature review," Technical Report, OECD 2016.

Hanushek, Eric A, Paul E Peterson, Laura M Talpey, and Ludger Woessmann, "Long-run trends in the US SES-achievement gap," Technical Report, National Bureau of Economic Research 2020.

Hoxby, Caroline M and Gretchen Weingarth, "Taking race out of the equation: School reassignment and the structure of peer effects," Technical Report, Citeseer 2005.

Kalogrides, Demetra, Susanna Loeb, and Tara Béteille, "Systematic sorting: Teacher characteristics and class assignments," *Sociology of Education*, 2013, *86* (2), 103–123.

Kane, Thomas J, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger, "Have we identified effective teachers? Validating measures of effective teaching using random assignment," in "Research Paper. MET Project. Bill & Melinda Gates Foundation" Citeseer 2013.

Kilpatrick, J., W.G. Martin, D. Schifter, and National Council of Teachers of Mathematics, *A Research Companion to Principles and Standards for School Mathematics*, National Council of Teachers of Mathematics, 2003.

Kilpatrick, Jeremy, Jane Swafford, Bradford Findell, and National Research Council, *Adding It Up: Helping Children Learn Mathematics*, Washington, DC: The National Academies Press, 2001.

Kremer, Michael, Conner Brannen, and Rachel Glennerster, "The challenge of education and learning in the developing world," *Science*, 2013, *340* (6130), 297–300.

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

Macartney, Hugh, Robert McMillan, and Uros Petronijevic, "A Quantitative Framework for Analyzing the Distributional Effects of Incentive Schemes," Working Paper 28816, National Bureau of Economic Research 2021.

Manski, Charles F, "Identification of endogenous social effects: The reflection problem," *The review of economic studies*, 1993, *60* (3), 531–542.

Neal, Derek and Diane Whitmore Schanzenbach, "Left behind by design: Proficiency counts and test-based accountability," *The Review of Economics and Statistics*, 2010, *92* (2), 263–283.

Polikoff, Morgan S and Andrew C Porter, "Instructional alignment as a measure of teaching quality," *Educational Evaluation and Policy Analysis*, 2014, *36* (4), 399–416.

_ , _ , and John Smithson, "How well aligned are state assessments of student achievement with state content standards?," *American Educational Research Journal*, 2011, *48* (4), 965–995.

Porter, Andrew C and John L Smithson, "Defining, Developing, and Using Curriculum Indicators. CPRE Research Report Series.," Technical Report 2001.

Reback, Randall, "Teaching to the rating: School accountability and the distribution of student achievement," *Journal of Public Economics*, 2008, *92* (5-6), 1394–1415.

Rivkin, Steven G, Eric A Hanushek, and John F Kain, "Teachers, schools, and academic achievement," *Econometrica*, 2005, *73* (2), 417–458.

Sacerdote, Bruce, "Peer effects in education: How might they work, how big are they and how much do we know thus far?," in "Handbook of the Economics of Education," Vol. 3, Elsevier, 2011, pp. 249–277.

Springer, Matthew G, *New York City's school-wide bonus pay program: Early evidence from a randomized trial*, DIANE Publishing, 2011.

Todd, Petra and Kenneth I Wolpin, "On the specification and estimation of the production function for cognitive achievement," *The Economic Journal*, 2003, *113* (485), F3–F33.

_ **and** _ , "The production of cognitive achievement in children: Home, school, and racial test score gaps," *Journal of Human capital*, 2007, *1* (1), 91–136.

_ **and** _ , "Accounting for Mathematics Performance of High School Students in Mexico: Estimating a Coordination Game in the Classroom," *Journal of Political Economy*, 2018, *126* (6), 2608–2650.

Table 1: Student and Teacher Characteristics

	Mean	Std.Dev		Mean
Panel A: Students (Obs. = 2532)				
Age	9.52	0.50	Gifted	0.06
Male	0.48		Special education (SpEd)	0.09
White	0.26		English language learner (ELL)	0.17
Black	0.43		Reduced price/free lunch	0.45
Hispanic	0.25			
N. books in bedroom:			N. computers at home:	
None	0.09		None	0.12
≥1 and ≤10	0.22		One	0.45
≥11 and ≤24	0.21		More than one	0.43
≥25	0.48			
Has person at home to help with homework:			Has no quiet place to study at home:	
Never	0.02		Never	0.40
Mostly not	0.03		Mostly not	0.12
Sometimes	0.09		Sometimes	0.16
Mostly	0.17		Mostly	0.12
Always	0.69		Always	0.2
Panel B: Teachers (Obs. = 101)				
Years of experience in the district	6.40	5.94		
Master's degree	0.53			
Teaches both Math and ELA (generalist)	0.83			

Table 2: Class Time Allocation, State Curriculum Standards, and Classroom Composition

	Mean	St.Dev	p25	Median	p75
Panel A: Class time allocation (% of total class time)					
Place value, rounding, addition, and subtraction	6.797	2.165	5.614	6.678	7.567
Multi-digit multiplication and division	25.981	6.246	21.967	25.229	28.899
Shapes, angles, and geometry	17.513	5.135	14.014	17.448	21.078
Fractions and decimals	25.546	5.367	22.362	25.464	29.427
Unit conversion and measurement	24.163	6.182	20.472	24.627	28.352
Panel B: Curriculum standards (% of test content)					
Place value, rounding, addition, and subtraction	7.967	2.632			
Multi-digit multiplication and division	29.483	6.023			
Shapes, angles, and geometry	15.910	5.157			
Fractions and decimals	25.617	9.222			
Unit conversion and measurement	21.024	7.851			
Panel C: Distance between instruction and standards (abs. dev.)					
Place value, rounding, addition, and subtraction	2.471	2.201			
Multi-digit multiplication and division	7.601	5.152			
Shapes, angles, and geometry	5.887	4.179			
Fractions and decimals	9.288	5.376			
Unit conversion and measurement	8.629	5.338			

Notes: Curriculum standards for each topic (Panel B) are measured as the percentage of items in the 4^{th} grade state test in mathematics related to each different topic. Information on the test content has been collected by the MET staff and categorized to be consistent with the topic categories included in the SEC survey. Panel C shows statistics of the absolute deviation between observed class time allocation choices and the standards.

Table 3: Classroom Composition and Tracking Intensity

	Mean	St.Dev	p25	Median	p75
Student composition across classrooms (baseline knowledge):					
% of students in 1^{st} tercile	30.20	21.80	14.30	26.30	46.40
% of students in 2^{nd} tercile	33.20	13.20	24.10	34.80	41.70
% of students in 3^{rd} tercile	36.60	26.80	14.30	30.20	52.00
Class size	23.29	4.97	20	23	26
N. classrooms	101				
Share of between-classroom variation in baseline knowledge:					
(based on all Year-1 MET schools)					
Data (1)	0.074	0.089	0.016	0.039	0.099
Tracking (2)	0.616	0.158	0.521	0.626	0.729
Tracking Intensity (ratio of (1) to (2))	0.123	0.158	0.027	0.062	0.145
N. schools	111				

Notes: The upper panel reports statistics on the distribution of the percentages of students in each district-level tercile of baseline knowledge (based on the 3^{rd} grade state tests in math) across classrooms in the sample. The lower panel shows the degree at which schools participating in the first year of the MET study are tracking 4^{th} grade students in math classrooms. Let $Var_s(K_0) = Var_s(K_0)^B + Var_s(K_0)^W$ denote the decomposition of the total variance of baseline knowledge in school s, $Var_s(K_0)$, in between-classroom ($Var_s(K_0)^B$) and within-classroom ($Var_s(K_0)^W$) variation. The share of between-classroom variation is then $Var_s(K_0)^B / Var_s(K_0)$. The second row of the lower panel (*Tracking (2)*) shows the share of between-classroom variation in the most extreme version of tracking (i.e., students ranked from low to high based on baseline knowledge and then reassigned sequentially to each classrooms, keeping equal class sizes in each school). The ratio of the two shares (third row) gives the measure of *tracking intensity* in each school.

Table 4: Descriptive Statistics of the Latent Factors Measures

Student knowledge:	Mean	Std.Dev			
3^{rd} grade math state test score (rescaled)	510.179	95.415			
BAM test (% correct answers)	54.735	21.389			
Teacher ability					
CLASS Behavior management (1-7 scale)	5.943	0.715			
CLASS Content understanding (1-7 scale)	4.137	0.481			
CLASS Productivity (1-7 scale)	5.918	0.555			
FFTM Management of class procedures (1-4 scale)	2.763	0.354			
FFTM Management of student behavior (1-4 scale)	2.840	0.344			
MQI Richness of mathematics (1-3 scale)	1.340	0.261			
MQI Mathematical knowledge for teaching (MKT) score (1-3 scale)	2.030	0.218			
Teacher explains clearly (0-4 scale)	3.321	0.295			
Teacher controls class behavior (0-4 scale)	2.251	0.437			
Teacher explains in orderly way (0-4 scale)	3.180	0.300			
Teacher can explain in several ways (0-4 scale)	3.216	0.295			
Teacher effort (0-4 scale)					
Teacher explains in another way if we do not understand	3.325	0.285			
Teacher pushes us to work hard	3.092	0.370			
Teacher does not waste time in class	2.664	0.385			
Teacher asks us if we understand the lesson	3.329	0.315			
Teacher asks us if we are following along	3.440	0.277			
Teacher writes feedback on our papers	2.887	0.387			
Teacher takes the time to summarize the lesson	2.813	0.480			
Teacher encourage us to do our best	3.533	0.257			
Teacher preference for adherence to standards (0-4 scale)					
Administrators require rigid adherence to standards	3.137	0.800			
I frequently refer to and use information found in standards documents	2.422	0.521			
Student effort	Never	Mostly not	Sometimes	Mostly	Always
I have done my best quality work in this class	0.007	0.010	0.089	0.242	0.457
In this class, I stop trying when the work gets hard	0.488	0.119	0.103	0.046	0.048
In this class, I take it easy and do not try to do my best	0.427	0.096	0.090	0.066	0.119
	None	Some	Most	All	All plus extra
How much homework do you usually complete?	0.006	0.062	0.106	0.489	0.137
Student preference for own knowledge	Never	Mostly not	Sometimes	Mostly	Always
School work is interesting	0.062	0.084	0.285	0.255	0.315
School work is not very enjoyable	0.287	0.165	0.256	0.124	0.168
I read at home almost every day	0.056	0.081	0.219	0.228	0.417

Notes: The 3^{rd} grade state test scores is rescaled to display mean and standard deviation equal to 500 and 100. The first 7 measures of teaching ability are based on several evaluation protocols of video-recorded lessons performed by the MET staff. All measures of teacher's instructional effort and the last 4 of teaching ability are average scores from student survey responses, whose categories ranged from 0 ("Never") to 4 ("Always").

Table 5: Production and Utility Functions Parameter Estimates

Panel A: Selected Production Function Parameters

Parameter	Label	Value (1)	Std.Err. (2)	Parameter	Label	Value (3)	Std.Err. (4)
δ_0	deprec. rate/unit conv.	0.1104	0.0034	π_{11}	direct peer spillov. $q=1$, $k=1$	-0.0236	0.0248
δ_1	unit conversion (mean)	0.0005	-	π_{13}	direct peer spillov. $q=1$, $k=3$	-0.0315	0.0316
γ_0	elasticity baseline knowledge	1.1288	0.0341	π_{21}	direct peer spillov. $q=2$, $k=1$	-0.0236	0.0316
γ_1	elasticity teacher ability	0.3414	0.0359	π_{23}	direct peer spillov. $q=2$, $k=3$	0.0607	0.0280
γ_{20}	elasticity teach. eff. (const.)	0.0231	0.0024	π_{31}	direct peer spillov. $q=3$, $k=1$	0.0861	0.0447
γ_{21}	elast. teach. eff. (class size)	-0.0004	0.0001	π_{33}	direct peer spillov. $q=3$, $k=3$	0.1033	0.0266
γ_3	elasticity student effort	0.0066	0.0067				

Panel B: Elasticity of Class Time Inputs

		Student tercile (q)					
		$q=1$		$q=2$		$q=3$	
Parameter	Topic group	Value (1)	Std.Err. (2)	Value (3)	Std.Err. (4)	Value (5)	Std.Err. (6)
η_{1q}	Place value, rounding, addition, and subtraction	0.0475	0.0068	0.0731	0.0111	0.0637	0.0073
η_{2q}	Multi-digit multiplication and division	0.2500	0.0149	0.2472	0.0271	0.2697	0.0196
η_{3q}	Shapes, angles, and geometry	0.0149	0.1096	0.0537	0.1824	0.0703	0.0393
η_{4q}	Fractions and decimals	0.3846	0.0237	0.2697	0.0196	0.3390	0.0166
η_{5q}	Unit conversion and measurement	0.3030	-	0.3563	-	0.2573	-

Panel C: Teacher Utility Parameters

Parameter	Label	Value (1)	Std.Err. (2)	Parameter	Label	Value (3)	Std.Err. (4)
ω_1^1	weight on 1st tercile student	29.3623	0.0269	$\widetilde{\alpha}_{13}$	preference for topic 3	0.3825	0.0654
ω_1^2	weight on 2nd tercile student	21.7276	0.1431	$\widetilde{\alpha}_{14}$	preference for topic 4	-0.0579	0.0464
ω_1^3	weight on 3rd tercile student	13.0701	0.5923	α_{201}	adherence to stand. topic 1	$5.9E{-}06$	0.0025
ω_{21}	weight on female student	-2.3023	0.0923	α_{202}	adherence to stand. topic 2	$1.4E{-}05$	0.0015
ω_{22}	weight on black student	4.8183	0.1027	α_{203}	adherence to stand. topic 3	0.0675	0.0041
ω_{23}	weight on Hispanic student	7.7529	0.5156	α_{204}	adherence to stand. topic 4	$5.3E{-}05$	0.0016
$\widetilde{\alpha}_{11}$	preference for topic 1	0.0891	0.0480	α_{205}	adherence to stand. topic 5	0.0002	0.0020
$\widetilde{\alpha}_{12}$	preference for topic 2	0.0712	0.0450	α_{21}	adherence to stand. (slope)	$7.6E{-}06$	0.0010

Notes: The table reports the parameter estimates of the production function and teacher utility. In Panel A, the value of δ_1 is the mean of the district-level parameters $(\delta_{1d})_{d=1}^5$ (whose estimates and standard errors are not reported due to confidentiality). For the parameters π_{qk}, the subscript q represents the student's own tercile while k is the peers' tercile. Both measures of K_{0ti} and K_{1ti} have been divided by 100 before the estimation. Hence, all parameters have to be interpreted accordingly. In Panel B, the elasticities of time spent teaching unit conversion and measurement are computed as $\eta_{5q} = 1 - \sum_{j=1}^4 \eta_{jq}$. In Panel C, the topic numbers from 1 to 5 refer to the same order of the topic groups in Panel B (e.g., topic 1 = "Place value, rounding, addit. and subract.", topic 2 = "Multi-digit multiplication and division", etc.). The parameters $(\widetilde{\alpha}_{11}, \widetilde{\alpha}_{12}, \widetilde{\alpha}_{13}, \widetilde{\alpha}_{14})$ are defined as $\widetilde{\alpha}_{1j} = \alpha_{1j} - \alpha_{15}$, $j = 1, \ldots, 4$.

Table 6: Within-Sample Model Fit

	Data		Model		
	Mean	Std.Dev.	Mean	Std.Dev.	$\sigma_{\text{true}}/\sigma_{\text{total}}$
	(1)	(2)	(3)	(4)	(5)
Knowledge measures:					
3^{rd} grade math state test score	510.179	95.415	515.881	97.500	
BAM test score (% correct)	54.735	21.389	51.810	20.714	0.531
Class time topic area (% of total class time):					
Place value, rounding, addition, and subtraction	6.797	2.165	7.721	6.071	
Multi-digit multiplication and division	25.981	6.246	26.080	6.650	
Shapes, angles, and geometry	17.513	5.135	17.244	6.674	
Fractions and decimals	25.546	5.367	25.731	7.076	
Unit conversions and measurement	24.163	6.182	23.224	7.122	
Teacher effort					
Teacher explains in another way if class does not understand	3.325	0.285	3.330	0.269	0.155
Teacher pushes students to work hard	3.092	0.370	3.108	0.349	0.003
Teacher does not waste time	2.664	0.385	2.682	0.375	0.146
Teacher asks questions to make sure students understand	3.329	0.315	3.364	0.300	0.436
Teacher asks if students are following along	3.440	0.277	3.445	0.288	0.377
Teacher writes feedback on our papers	2.887	0.387	2.897	0.418	0.077
Teacher takes time to summarize the lesson	2.813	0.480	2.819	0.442	0.142
Teacher encourages students to do their best	3.533	0.257	3.571	0.263	0.145
Student effort					
I have done my best quality work in this class	3.406	0.801	3.387	0.810	0.109
In this class, I stop trying when the work gets hard	0.817	1.214	0.854	1.240	0.141
In this class, I take it easy and do not try to do my best	1.190	1.512	1.236	1.531	0.102
How much homework do you usually complete?	3.862	0.814	3.847	0.820	0.104
Teacher ability					
CLASS Behavior management scale	5.943	0.715	5.917	0.761	0.574
CLASS Content understanding scale	4.137	0.481	4.153	0.502	0.307
CLASS Productivity scale	5.918	0.555	5.936	0.633	0.473
FFTM Management of class procedures score	2.763	0.354	2.727	0.442	0.806
FFTM Management of student behavior score	2.840	0.344	2.822	0.391	0.791
MQI Richness of mathematics score	1.340	0.261	1.338	0.250	0.011
MQI Mathematical knowledge for teaching (MKT) score	2.030	0.218	2.016	0.268	0.136
Teacher explains clearly (0-4 score)	3.321	0.295	3.319	0.301	0.222
Teacher controls class behavior (0-4 score)	2.251	0.437	2.256	0.443	0.275
Teacher explains in orderly way (0-4 score)	3.180	0.300	3.196	0.298	0.198
Teacher can explain in several ways (0-4 score)	3.216	0.295	3.209	0.278	0.326
Student preference for own knowledge					
I read at home almost every day	2.869	1.202	2.869	1.181	0.281
School work is interesting	2.677	1.178	2.647	2.676	0.085
School work is not very enjoyable	2.278	1.426	2.195	2.278	0.070
Teacher preference for adherence to standards					
Administrators require rigid adherence to standards	3.137	0.800	3.013	0.772	0.239
I frequently refer to and use info. found in stand. documents	2.422	0.521	2.366	0.509	0.827

Table 7: Out-of-Sample Validation (Year 2 Data Sample Fit)

	Data		Model	
	Mean	Std.Dev.	Mean	Std.Dev.
	(1)	(2)	(3)	(4)
Knowledge measures:				
BAM test score (% correct)	53.453	22.123	50.488	20.062
Class time topic area (% of total class time):				
Place value, rounding, addition, and subtraction			7.376	3.745
Multi-digit multiplication and division			26.562	6.289
Shapes, angles, and geometry			15.268	6.790
Fractions and decimals			27.926	6.726
Unit conversions and measurement			22.868	5.741
Teacher effort				
Teacher explains in another way if class does not understand	3.328	0.299	3.389	0.277
Teacher pushes students to work hard	3.192	0.427	3.107	0.369
Teacher does not waste time	2.709	0.402	2.744	0.404
Teacher asks questions to make sure students understand	3.392	0.322	3.418	0.329
Teacher asks if students are following along	3.502	0.252	3.528	0.300
Teacher writes feedback on paper	2.959	0.468	2.913	0.403
Teacher takes time to summarize lesson	2.981	0.434	2.894	0.472
Teacher encourages students to do their best	3.600	0.289	3.589	0.276
Teacher ability				
CLASS Behavior management scale	5.803	0.512	5.886	0.731
CLASS Content understanding scale	4.120	0.496	4.133	0.509
CLASS Productivity scale	5.803	0.419	5.901	0.585
FFTM Management of class procedures score	2.691	0.346	2.734	0.440
FFTM Management of student behavior score	2.767	0.380	2.820	0.398
MQI Richness of mathematics score	1.353	0.263	1.310	0.235
MQI Mathematical knowledge for teaching (MKT) score	2.027	0.225	2.002	0.249
Teacher explains clearly (0-4 score)	3.324	0.269	3.320	0.291
Teacher controls class behavior (0-4 score)	2.211	0.506	2.230	0.447
Teacher explains in orderly way (0-4 score)	3.229	0.347	3.187	0.301
Teacher can explain in several ways (0-4 score)	3.311	0.292	3.212	0.304
Student effort				
I have done my best quality work in this class	3.409	0.815	3.403	0.803
In this class, I stop trying when the work gets hard	0.831	1.246	0.801	1.205
In this class, I take it easy and do not try to do my best	1.316	1.519	1.166	1.507
How much homework do you usually complete?	3.933	0.833	3.863	0.820
Student preference for own knowledge				
School work is interesting	2.830	1.117	2.629	1.182
School work is not very enjoyable	1.596	1.398	2.257	1.431
Teacher preference for adherence to standards				
Administrators require rigid adherence to standards	2.865	0.870	1.539	0.668
I frequently refer to and use info. found in stand. documents	2.341	0.631	0.945	0.227

Table 8: Impact of Ability Tracking on Class Time Allocation

	Average % of Total Class Time Allocated		Tailored
	Baseline	Tracking	Instruction ($\widetilde{\tau}^q$)
	(1)	(2)	(3)
<u>1^{st} tercile students</u>			
Place value, round., addit., and subtr.	7.032	6.123	4.747
Multi-digit multipl. and division	25.435	24.799	24.997
Shapes, angles, and geom.	17.402	16.524	1.486
Fractions and decimals	26.825	27.804	38.458
Unit conversion and meas.	23.305	24.750	30.312
<u>2^{nd} tercile students</u>			
Place value, round., addit., and subtr.	7.977	8.481	7.314
Multi-digit multipl. and division	26.112	25.738	24.722
Shapes, angles, and geom.	15.981	16.215	5.367
Fractions and decimals	27.017	27.383	37.848
Unit conversion and meas.	22.913	22.183	24.749
<u>3^{rd} tercile students</u>			
Place value, round., addit., and subtr.	8.269	8.753	6.374
Multi-digit multipl. and division	26.735	28.049	26.965
Shapes, angles, and geom.	14.980	15.529	7.034
Fractions and decimals	26.778	25.141	33.900
Unit conversion and meas.	23.238	22.528	25.727

Notes: The baseline levels of class time allocation are simulated using the student-classroom assignment observed in the data. Counterfactual results are, instead, based on tracking with random assignment of teachers to classrooms. The level of tailored instruction for students in tercile $q = 1,2,3$ is given by $\widetilde{\boldsymbol{\tau}}^q = \overline{\tau} \times \boldsymbol{\eta}_q = 100 \times (\eta_{1q}, \eta_{2q}, \eta_{3q}, \eta_{4q}, \eta_{5q})$ as illustrated in Section 2.2.

Table 9: Disentangling Peer-to-Peer Spillovers

	ΔBAM score (in SD)			
	All terciles	1^{st} tercile	2^{nd} tercile	3^{rd} tercile
Tracking with random assignment of teachers				
Total effect	0.0209	−0.0075	−0.0358	0.1091
Direct peer spillovers	0.0205	−0.0127	−0.0333	0.1103
Indirect peer spillovers	0.0004	0.0052	−0.0025	−0.0012
Tracking with positive assortative matching				
Total effect	0.0264	−0.0383	−0.0368	0.1574
Direct peer spillovers	0.0205	−0.0127	−0.0333	0.1103
Indirect peer spillovers	0.0059	−0.0256	−0.0035	0.0471
Tracking with negative assortative matching				
Total effect	0.0148	0.0266	−0.0318	0.0525
Direct peer spillovers	0.0205	−0.0127	−0.0333	0.1103
Indirect peer spillovers	−0.0057	0.0393	0.0015	−0.0578

Notes: The table shows the decomposition of the effect of tracking on students end-of-year knowledge (measured by the standardized BAM score) at different terciles. Direct peer spillovers represent the effect of tracking due to classmates characteristics when teachers do not adjust instruction to the new composition of the classroom (i.e., each student receives the same instruction as the baseline "non-tracking" scenario). Indirect peer spillovers represent the effect of tracking due to teachers instructional response to the new classroom composition, while keeping direct peer spillovers fixed to the baseline level. Teacher assortative matching is based on teaching ability.

Table 10: Impact of Teaching According to the Curriculum Standards on End-of-Year Knowledge

	BAM score (% correct)		
Baseline knowledge tercile	Baseline	Counterf.	Δ (in SD)
All terciles (SD = 20.714)	51.8104	51.7459	−0.0031
1^{st} tercile	35.1232	35.0190	−0.0050
2^{nd} tercile	51.3147	51.2401	−0.0036
3^{rd} tercile	69.3925	69.3791	−0.0006

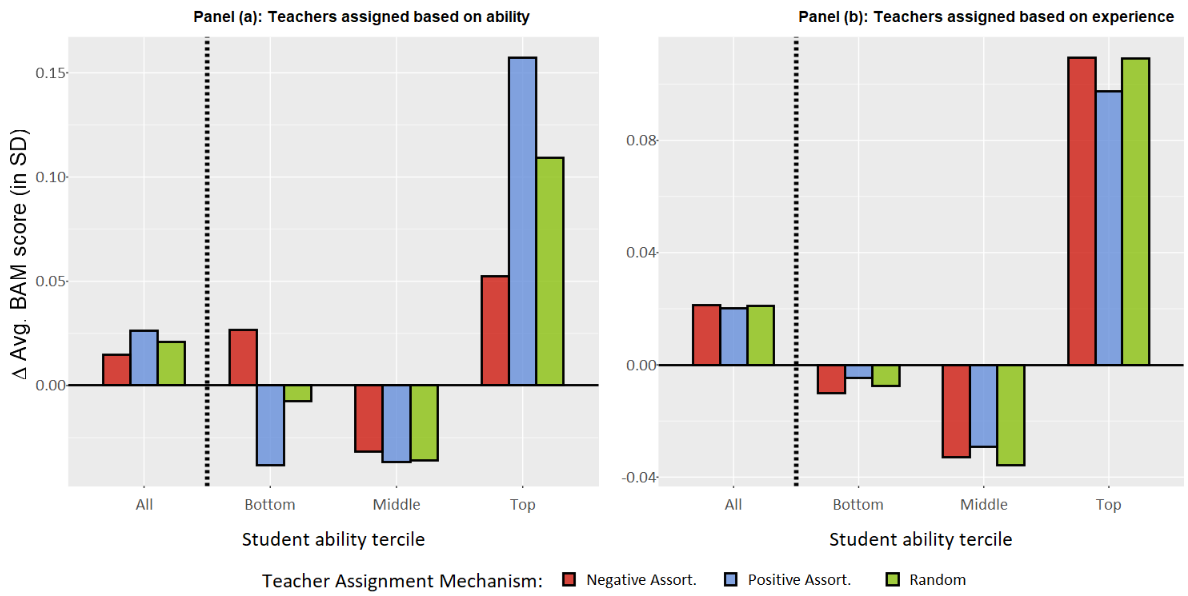

Figure 1: Impact of Ability Tracking on Student Achievement

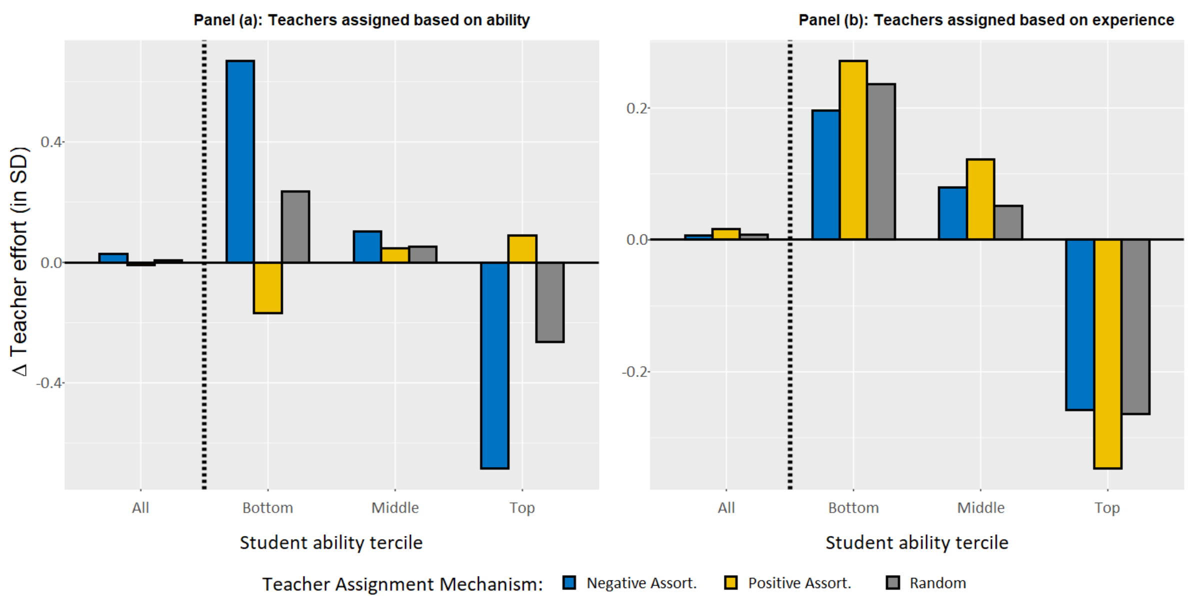

Figure 2: Impact of Ability Tracking on Teacher Effort

Appendix

A. Model Solution and Likelihood Function

Let $D_{ti} \equiv \delta_1 K_{0ti}^{\gamma_0} A_t^{\gamma_1}$ and $F_q(\boldsymbol{\tau}_t) \equiv \prod_{j=1}^{J} \tau_{tj}^{\eta_{jq}}$, and also $\tilde{\alpha}_{1k} \equiv \alpha_{1k} - \alpha_{1J}$ and $\tilde{\varepsilon}_{tk} \equiv \varepsilon_{tk} - \varepsilon_{tJ}$. The FOCs (7a), (7b), and (9) can be re-written as

$$e_t^* = \gamma_2^{\frac{1}{2-\gamma_2}} \gamma_3^{\frac{\gamma_3}{(2-\gamma_3)(2-\gamma_2)}} \left[\sum_{\ell=1}^{N_t} \omega_{t\ell} [D_{t\ell} F_{q_\ell}(\boldsymbol{\tau}_t^*)]^{\frac{2}{2-\gamma_3}} (\psi_{t\ell})^{\frac{\gamma_3}{2-\gamma_3}} \right]^{\frac{1}{2(2-\gamma_2-\gamma_3)}}$$

$$h_{ti}^* = \gamma_2^{\frac{\gamma_2}{2(2-\gamma_2-\gamma_3)}} \gamma_3^{\frac{2-\gamma_2}{2(2-\gamma_2-\gamma_3)}} \left[\psi_{ti} D_{ti} F_q(\boldsymbol{\tau}_t^*)\right]^{\frac{1}{2-\gamma_3}} \left[\sum_{\ell=1}^{N_t} \omega_{t\ell} [D_{t\ell} F_{q_\ell}(\boldsymbol{\tau}_t^*)]^{\frac{2}{2-\gamma_3}} (\psi_{t\ell})^{\frac{\gamma_3}{2-\gamma_3}} \right]^{\frac{\gamma_2}{2(2-\gamma_2-\gamma_3)}}$$

and, for $k = 1, \ldots, J-1$,

$$-\tilde{\varepsilon}_{tk} = \sum_{i=1}^{N_t} \omega_{ti} D_{ti} e_{st}^{*\gamma_2} h_{ti}^{*\gamma_3} (\eta_{kq_i} \tau_{tk}^{*-1} - \eta_{Jq} \tau_{tJ}^{*-1}) F_q(\boldsymbol{\tau}_t^*) \tag{13}$$

$$\tilde{\alpha}_{1k} - (\alpha_2^k + \alpha_{21}\phi_t)\left(\tau_{tk}^* - \varphi_{tk}\right) + (\alpha_2^J + \alpha_{21}\phi_t)\left(\tau_{tJ}^* - \varphi_{tJ}\right) \tag{14}$$

For each student i taught by teacher t, an observation in the data includes the measures of both exogenous and endogenous latent factors together with the initial conditions,

$$O_{ti} = \left((A_t^m)_{m=1}^{M_A}, (e_t^m)_{m=1}^{M_e}, (\phi_t^m)_{m=1}^{M_\phi}, (h_{ti}^m)_{m=1}^{M_h}, (\psi_{ti}^m)_{m=1}^{M_\psi}, K_{0ti}, K_{1ti}, \boldsymbol{\tau}_t^*, X_{ti} \right),$$

with $X_{ti} = (X_t^A, X_t^\phi, X_{ti}^{K_0}, X_{ti}^\psi, \varphi_t, W_{ti})$. The vectors of observations and initial conditions for class t are then $\boldsymbol{O}_t = (O_{t1}, \ldots, O_{tN_t})$ and $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{tN_t})$. Define the vectors of random effects $\boldsymbol{\chi}_t = \left(\boldsymbol{v}_t, \boldsymbol{\zeta}_{t1}, \ldots, \boldsymbol{\zeta}_{tN_t}\right)$. In order to derive the likelihood contribution of class t, let's first assume that $\boldsymbol{\chi}_t$ is observed by the econometrician. Then, given the distributional assumptions and suppressing the subscripts to ease notation, the likelihood of the exogenous latent factor $\theta \in \{A, \phi, \psi\}$ is given by

$$\ell_\theta\left((\theta^m)_{m=1}^{M_\theta} | X^\theta, \boldsymbol{\chi}\right) = \prod_{m=1}^{M_\theta} \ell_\theta^m(\theta^m | \theta)$$

where $\ell_\theta^m(\cdot)$ is the likelihood of the m^{th} measure of θ. As for the endogenous variables, conditional on $\left(\boldsymbol{X}_t, \boldsymbol{K}_{0t}, \boldsymbol{\tau}_t, \boldsymbol{\chi}_t\right)$, optimal effort e_t^* and h_{ti}^* are completely deterministic. Hence, the likelihood of the effort measures are given by

$$\ell_e\big((e_t^m)_{m=1}^{M_e}|\underline{X}_t,\boldsymbol{\tau}_t,\boldsymbol{\chi}_t\big)=\prod_{m=1}^{M_e}\ell_e^m\big(e_t^m|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\boldsymbol{\chi}_t\big),$$
$$\ell_h\big((h_{ti}^m)_{m=1}^{M_h}|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\boldsymbol{\chi}_t\big)=\prod_{m=1}^{M_h}\ell_h^m\big(h_t^m|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\boldsymbol{\chi}_t\big)$$

.

Similarly, the conditional likelihood of end-of-year knowledge K_{1ti} is given by $\ell_{K_1}\big(K_{1ti}|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\boldsymbol{\chi}_t\big)$. Finally, the conditional likelihood of $\boldsymbol{\tau}$ is derived from the FOCs (13), for $k=1\ldots,J-1$, combined with the distributional assumption on $\boldsymbol{\varepsilon}_t$. Specifically, denoting the RHS of (13) for all $k=1\ldots,J-1$ as a multivariate function $\widetilde{\boldsymbol{\varepsilon}}(\boldsymbol{X}_t,\boldsymbol{\chi}_t)$ we have that the likelihood of $\boldsymbol{\tau}$ is given by

$$\ell_{\boldsymbol{\tau}}\big(\boldsymbol{\tau}_t|\boldsymbol{X}_t,\boldsymbol{\chi}_t\big)=\Big|\det J\big(\widetilde{\boldsymbol{\varepsilon}}(\boldsymbol{X}_t,\boldsymbol{\chi}_t)\big)\Big|\times\ell_{\widetilde{\boldsymbol{\varepsilon}}}\big(\widetilde{\boldsymbol{\varepsilon}}(\boldsymbol{X}_t,\boldsymbol{\chi}_t)|\boldsymbol{X}_t,\boldsymbol{\chi}_t\big)$$

where $J\big(\widetilde{\boldsymbol{\varepsilon}}(\boldsymbol{X}_t,\boldsymbol{\chi}_t)\big)$ is the Jacobian matrix of $\widetilde{\boldsymbol{\varepsilon}}\big(\boldsymbol{X}_t,\boldsymbol{\chi}_t\big)$ (with derivatives with respect to $(\tau_{t1},\ldots,\tau_{tJ-1})$) and $\ell_{\widetilde{\boldsymbol{\varepsilon}}}(\cdot)$ the likelihood of $\widetilde{\boldsymbol{\varepsilon}}_t$ (a multivariate normal with mean zero and covariance matrix $\Sigma_{\widetilde{\boldsymbol{\varepsilon}}}$).

As a result, the likelihood contribution of class t in school s conditional on $\big(\boldsymbol{X}_t,\boldsymbol{\chi}_t\big)$ is given by

$$\begin{aligned}
&L_t\big(\Theta\big|\boldsymbol{O}_t,\boldsymbol{\chi}_t\big)=\\
&=\prod_{i=1}^{N_t}\Big[\ell_{K_1}\big(K_{1ti}|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\boldsymbol{\chi}_t\big)\ell_h\big((h_{ti}^m)_{m=1}^{M_h}|\boldsymbol{X}_t,\boldsymbol{\chi}_t\big)\ell_\psi\big((\psi_{ti}^m)_{m=1}^{M_\psi}|X_{ti}^\psi,\boldsymbol{\chi}_t\big)\ell_{K_0}\big(K_{0ti}|X_{ti}^{K_0},\boldsymbol{\chi}_t\big)\boldsymbol{\Phi}\big(\boldsymbol{\zeta}_{ti}|\boldsymbol{X}_t,\boldsymbol{\upsilon}_t;\Sigma_\zeta\big)\Big]\times\\
&\times\ell_e\big((e_t^m)_{m=1}^{M_e}|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\boldsymbol{\chi}_t\big)\ell_A\big((A_t^m)_{m=1}^{M_A}|X_t^A,\boldsymbol{\chi}_t\big)\ell_\phi\big((\phi_t^m)_{m=1}^{M_\phi}|X_t^\phi,\boldsymbol{\chi}_t\big)\times\ell_{\boldsymbol{\tau}}\big(\boldsymbol{\tau}_t|\boldsymbol{X}_t,\boldsymbol{\chi}_t\big)\boldsymbol{\Phi}\big(\boldsymbol{\upsilon}_t|\boldsymbol{X}_t;\Sigma_\upsilon\big)
\end{aligned}\tag{15}$$

with $\boldsymbol{\Phi}(\cdot;\Sigma)$ a multivariate normal density with zero mean and covariance matrix Σ. Now since $\boldsymbol{\chi}_t$ is actually unobserved, we need to integrate it out, that is

$$L_t(\Theta|\boldsymbol{O}_t)=\int L_t(\Theta|\boldsymbol{O}_t,\boldsymbol{\chi})d\boldsymbol{\chi}.\tag{16}$$

Given R draws from the joint distribution of $\boldsymbol{\chi}_t$, denoted $(\hat{\boldsymbol{\chi}}_{tr})_{r=1}^R$, we can perform a Monte Carlo integration to approximate (16) and obtain our simulated likelihood contribution of class t

$$\begin{aligned}
\hat{L}_t=\frac{1}{R}\sum_{r=1}^R\Bigg\{&\Bigg[\prod_{i=1}^{N_t}\ell_{K_1}\big(K_{1ti}|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\hat{\boldsymbol{\chi}}_t\big)\ell_h\big((h_{ti}^m)_{m=1}^{M_h}|\boldsymbol{X}_t,\hat{\boldsymbol{\chi}}_{tr}\big)\ell_\psi\big((\psi_{ti}^m)_{m=1}^{M_\psi}|X_{ti}^\psi,\hat{\boldsymbol{\chi}}_t\big)\ell_{K_0}\big(K_{0ti}|X_{ti}^{K_0},\hat{\boldsymbol{\chi}}_{tr}\big)\Bigg]\times\\
&\times\ell_e\big((e_t^m)_{m=1}^{M_e}|\boldsymbol{X}_t,\boldsymbol{\tau}_t,\hat{\boldsymbol{\chi}}_{tr}\big)\ell_A\big((A_t^m)_{m=1}^{M_A}|X_t^A,\hat{\boldsymbol{\chi}}_{tr}\big)\ell_\phi\big((\phi_t^m)_{m=1}^{M_\phi}|X_t^\phi,\hat{\boldsymbol{\chi}}_{tr}\big)\times\ell_{\boldsymbol{\tau}}\big(\boldsymbol{\tau}_t|\boldsymbol{X}_t,\hat{\boldsymbol{\chi}}_{tr}\big)\Bigg\}
\end{aligned}\tag{17}$$

B. Additional Tables and Figures

Table B.1: Correlation Matrices: Latent Factors Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Teacher ability (Pearson's correlation coeff.)											
CLASS Behavior management (1)	1.000										
p-value											
CLASS Content understanding (2)	0.383	1.000									
p-value	0.001										
CLASS Productivity (3)	0.810	0.495	1.000								
p-value	0.000	0.000									
FFTM Management of class procedures (4)	0.591	0.381	0.486	1.000							
p-value	0.000	0.001	0.000								
FFTM Management of student behavior (5)	0.680	0.338	0.450	0.753	1.000						
p-value	0.000	0.004	0.000	0.000							
MQI Richness of mathematics (6)	0.307	0.413	0.348	0.197	0.161	1.000					
p-value	0.009	0.000	0.003	0.102	0.182						
MQI Mathematical knowledge for teaching (MKT) score (7)	0.321	0.321	0.283	0.332	0.326	0.347	1.000				
p-value	0.007	0.007	0.018	0.005	0.006	0.003					
Teacher explains clearly (8)	0.119	0.101	0.179	0.155	0.074	0.055	−0.108	1.000			
p-value	0.325	0.407	0.139	0.201	0.545	0.649	0.374				
Teacher controls class behavior (9)	0.332	0.291	0.427	0.306	0.292	0.110	0.170	0.341	1.000		
p-value	0.005	0.014	0.000	0.001	0.014	0.364	0.159	0.000			
Teacher explains in orderly way (10)	0.199	0.061	0.199	0.025	0.039	0.127	0.007	0.587	0.325	1.000	
p-value	0.098	0.616	0.099	0.838	0.747	0.294	0.955	0.000	0.001		
Teacher can explain in several ways (11)	0.173	0.255	0.353	0.191	0.156	0.180	0.062	0.564	0.385	0.557	1.000
p-value	0.153	0.033	0.003	0.113	0.198	0.135	0.609	0.000	0.000	0.000	
Teacher effort (Pearson's correlation coeff.)											
Teacher explains in another way if we do not understand (1)	1.000										
p-value											
Teacher pushes us to work hard (2)	0.207	1.000									
p-value	0.041										
Teacher does not waste time in class (3)	0.150	0.319	1.000								
*p-value*e	0.140	0.001									
Teacher asks us if we understand the lesson (4)	0.365	0.360	0.167	1.000							
p-value	0.000	0.000	0.101								
Teacher asks us if we are following along (5)	0.409	0.380	0.016	0.609	1.000						
p-value	0.000	0.000	0.876	0.000							
Teacher writes feedback on our papers (6)	0.372	0.238	0.102	0.178	0.261	1.000					
p-value	0.000	0.018	0.319	0.079	0.009						
Teacher takes the time to summarize the lesson (7)	0.452	0.357	0.248	0.442	0.448	0.459	1.000				
p-value	0.000	0.000	0.014	0.000	0.000	0.000					
Teacher encourage us to do our best (8)	0.425	0.286	0.405	0.233	0.324	0.281	0.214	1.000			
p-value	0.000	0.004	0.000	0.021	0.001	0.005	0.035				
Teacher preference for adherence to standards (Pearson's χ^2)											
Administrators require rigid adherence to standards (1)	.										
p-value	.										
I frequently refer to and use information found in standards documents (2)	1.02	.									
p-value	0.60										
Student effort (Pearson's χ^2)											
I have done my best quality work in this class (1)	.										
p-value	.										
In this class, I stop trying when the work gets hard (2)	240.90	.									
p-value	0.00										
In this class, I take it easy and do not try to do my best (3)	276.08	393.21	.								
p-value	0.00	0.00									
How much homework do you usually complete? (4)	143.23	108.17	92.01	.							
p-value	0.00	0.00	0.00								
Student preference for own knowledge (Pearson's χ^2)											
School work is interesting (1)	.										
p-value	.										
School work is not very enjoyable (2)	512.35	.									
p-value	0.00										
I read at home almost every day (3)	228.46	95.70	.								
p-value	0.00	0.00									

Table B.2: Exogenous Inputs Equation Parameter Estimates

Determinants	Baseline knowledge (K_{0ti})		Student preference for own knowledge ($log(\psi_{ti})$)	
	Estimate	Std.Err.	Estimate	Std.Err.
Constant	5.8378	0.3880	0.0000	
Male	0.0032	0.0312	−0.2458	0.0243
Gifted	0.9097	0.0701	0.4049	0.0556
Special education	−0.4311	0.0579	−0.0800	0.0398
English language learner	−0.0065	0.0575	0.1430	0.0423
Free/Reduced price lunch	−0.1257	0.0385	0.0251	0.0273
Black	−0.4344	0.0527	−0.1867	0.0341
Hispanic	−0.2076	0.0571	−0.0360	0.0420
Age	−0.0853	0.0330	0.0398	0.0235
N. books in bedroom (base = None):				
≥1 and ≤10	0.0614	0.0595	0.3088	0.0420
≥11 and ≤24	0.1821	0.0608	0.3791	0.0431
≥25	0.1226	0.0560	0.5870	0.0432
Has quiet place to study at home:				
Mostly not	0.0945	0.0521	−0.2470	0.0424
Sometimes	−0.0370	0.0462	−0.2347	0.0362
Mostly	−0.1504	0.0528	−0.2822	0.0385
Always	−0.1654	0.0428	−0.0523	0.0299
N. computers at home (base = None)				
One	0.1075	0.0509	−0.0046	0.0370
More than one	0.2167	0.0544	−0.1288	0.0397
Has person at home to help with homework (base = Never)				
Mostly not	0.3237	0.1261	−0.1139	0.0886
Sometimes	0.3238	0.1087	0.0743	0.0759
Mostly	0.3694	0.1039	0.2241	0.0723
Always	0.2545	0.0987	0.4148	0.0700
	Teacher ability ($log(A_t)$)		Teacher preference for adherence to standards ($log(\psi_t)$)	
	Estimate	Std. Err.		
Constant	0.0000		0.0000	
Years of experience	0.0205	0.0261	−0.0077	0.0406
Years of experience (squared)	−0.0015	0.0014	0.0014	0.0023
Master's degree	0.1532	0.1305	0.5319	0.2288
Generalist teacher	−0.2222	0.2111	0.0465	0.3054

Table B.3: Exogenous Inputs Random Effects Covariances

Taste Shocks Covariance Matrix ($\Sigma_{\widetilde{\varepsilon}}$)

	$\widetilde{\varepsilon}_{11}$	$\widetilde{\varepsilon}_{12}$	$\widetilde{\varepsilon}_{13}$	$\widetilde{\varepsilon}_{14}$
$\widetilde{\varepsilon}_{11}$	0.0190			
	(0.0085)			
$\widetilde{\varepsilon}_{12}$	0.0147	0.0187		
	(0.0099)	(0.0063)		
$\widetilde{\varepsilon}_{13}$	−0.0013	−0.0024	0.1230	
	(0.0315)	(0.0350)	(0.0160)	
$\widetilde{\varepsilon}_{14}$	0.0109	0.0139	−0.0042	0.0251
	(0.0131)	(0.0129)	(0.0087)	(0.0092)

Teacher-level Random Effects Covariance Matrix (Σ_{ν})

	(1)	(2)	(3)	(4)
Teacher ability (1)	0.2922			
	(0.060)			
Teacher preference for adherence to standards (2)	−0.0311	0.0982		
	(0.0460)	(0.0939)		
Student preference for own knowledge (3)	0.0846	0.0793	0.1115	
	(0.2879)	(0.3471)	(0.0280)	
Baseline knowledge (4)	−0.0014	0.0186	−0.0074	0.0808
	(0.2179)	(0.2990)	(0.0413)	(0.0463)

Student-level Random Effects Covariance Matrix (Σ_{ζ})

	(1)	(2)
Student preference for own knowledge (1)	0.1228	
	(0.0223)	
Baseline knowledge (2)	0.0257	0.5079
	(0.0155)	(0.0088)

Notes: Standard errors in parenthesis

Table B.4: Measurement Equations Parameters - Endogenous Latent Factors

	Intercept (μ^{y}_{0m})		Slope (μ^{y}_{1m})		Meas.Error ($\sigma_{\varsigma ym}$)	
	Estimate	Std. Err.	Estimate	Std.Err.	Estimate	Std.Err.
End-of-year knowledge (K_{1ti})						
BAM test score (fraction correct)	·	·	1.000	·	0.1410	0.0023
Teacher effort (e_t)						
Teacher explains in another way if we do not understand	2.7028	0.0287	1.0000	0.0000	0.2620	0.0082
Teacher pushes us to work hard	2.9783	0.0365	0.1756	0.2369	0.3659	0.0155
Teacher does not waste time in class	1.8248	0.0378	1.3535	0.2330	0.3711	0.0192
Teacher asks us if we understand the lesson	2.1664	0.0311	1.8680	0.2141	0.2677	0.0094
Teacher asks us if we are following along	2.4006	0.0273	1.6690	0.1914	0.2272	0.0069
Teacher writes feedback on our papers	2.1979	0.0388	1.0965	0.2522	0.3843	0.0174
Teacher takes the time to summarize the lesson	1.8392	0.0492	1.5698	0.3443	0.4384	0.0226
Teacher encourage us to do our best	2.9471	0.0251	0.9463	0.1617	0.2437	0.0074
Student effort (h_{ti})						
I have done my best quality work in this class	0.0000		1.0000	0.0000	1.0000	
Cutoff 1 ("Never"-"Mostly not")	−1.6038	0.0804				
Cutoff 2 ("Mostly not"-"Sometimes")	−1.2247	0.0588				
Cutoff 3 ("Sometimes"-"Mostly")	−0.2623	0.0324				
Cutoff 4 ("Mostly"-"Always")	0.7383	0.0260				
In this class, I stop trying when the work gets hard	0.0000		1.1712	0.1164	1.0000	
Cutoff 1 ("Never"-"Mostly not")	−0.5988	0.0403				
Cutoff 2 ("Mostly not"-"Sometimes")	−0.2122	0.0330				
Cutoff 3 ("Sometimes"-"Mostly")	0.3271	0.0278				
Cutoff 4 ("Mostly"-"Always")	0.7803	0.0261				
In this class, I take it easy and do not try to do my best	0.0000		0.9598	0.0824	1.0000	
Cutoff 1 ("Never"-"Mostly not")	−0.2119	0.0317				
Cutoff 2 ("Mostly not"-"Sometimes")	0.1062	0.0285				
Cutoff 3 ("Sometimes"-"Mostly")	0.4576	0.0266				
Cutoff 4 ("Mostly"-"Always")	0.7900	0.0259				
How much homework do you usually complete?	0.0000		0.9735	0.0702	1.0000	
Cutoff 1 ("Never"-"Mostly not")	−1.6513	0.0835				
Cutoff 2 ("Mostly not"-"Sometimes")	−0.5495	0.0361				
Cutoff 3 ("Sometimes"-"Mostly")	0.0814	0.0287				
Cutoff 4 ("Mostly"-"Always")	1.8935	0.0306				

Notes: The BAM test score measure has been rescaled from “% correct” to “fraction” for computational purposes before the estimation (the measure K_{0ti} has been dived by 100 as well). The constant term of the measure of K_{1ti}, $\mu_0^{K_1}$, is allowed to vary by district. These values are not reported in the table due to a confidentiality agreement with ICPSR.

Table B.5: Measurement Equations Parameters - Exogenous Latent Factors

	Intercept (μ^y_{0m})		Slope (μ^y_{1m})		Meas.Error ($\sigma_{\varsigma ym}$)	
	Estimate	Std. Err.	Estimate	Std.Err.	Estimate	Std.Err.
Teacher ability ($log(A_t)$)						
CLASS Behavior management scale	6.1656	0.1892	1.0000	·	0.5042	0.0277
CLASS Content understanding scale	4.2827	0.0786	0.4818	0.0787	0.4240	0.0304
CLASS Productivity scale	6.1145	0.1349	0.7546	0.0699	0.4103	0.0183
FFTM Management of class procedures score	2.9415	0.0768	0.6927	0.0528	0.2303	0.0124
FFTM Management of student behavior score	2.9924	0.0722	0.6050	0.0507	0.2033	0.0115
MQI Richness of mathematics score	1.3651	0.0404	0.1630	0.0424	0.2590	0.0089
MQI Mathematical knowledge for teaching (MKT) score	2.0529	0.0316	0.1714	0.0330	0.2411	0.0054
Teacher explains clearly	3.3907	0.0383	0.2463	0.0486	0.2699	0.0120
Teacher controls class behavior	2.3622	0.0621	0.4032	0.0701	0.3892	0.0246
Teacher explains in orderly way	3.2438	0.0396	0.2299	0.0482	0.2755	0.0117
Teacher can explain in several ways	3.2922	0.0417	0.2749	0.0487	0.2634	0.0119
Student preference for own knowledge ($log(\psi_{ti})$)						
I read at home almost every day	0.0000		1.0000		1.0000	
Cutoff 1 ("Never"-"Mostly not")	−1.2339	0.5078				
Cutoff 2 ("Mostly not"-"Sometimes")	−0.6183	0.5072				
Cutoff 3 ("Sometimes"-"Mostly")	0.2833	0.5082				
Cutoff 4 ("Mostly"-"Always")	0.9971	0.5091				
School work is interesting	0.0000		0.4966	0.1858	1.0000	-
Cutoff 1 ("Never"-"Mostly not")	−1.2747	0.4847				
Cutoff 2 ("Mostly not"-"Sometimes")	−0.7550	0.4839				
Cutoff 3 ("Sometimes"-"Mostly")	0.1859	0.4834				
Cutoff 4 ("Mostly"-"Always")	0.8819	0.4841				
School work is not very enjoyable	0.0000		−0.4495	0.0930	1.0000	
Cutoff 1 ("Never"-"Mostly not")	−0.9216	0.4819				
Cutoff 2 ("Mostly not"-"Sometimes")	−0.4558	0.4815				
Cutoff 3 ("Sometimes"-"Mostly")	0.2457	0.4816				
Cutoff 4 ("Mostly"-"Always")	0.6773	0.4818				
Teacher preference for adherence to standards ($log(\phi_{ti})$)						
Administrators require rigid adherence to standards	0.0000		1.0000		1.0000	
Cutoff 2 ("Disagree"-"Neither")	−1.7035	0.5804				
Cutoff 3 ("Neither"-"Agree")	−0.7385	0.4501				
Cutoff 4 ("Agree"-"Strongly agree")	0.6952	0.4307				
I frequently refer to and use information found in standards documents	0.0000		4.3914	0.8928	1.0000	
Cutoff 3 ("Neither"-"Agree")	−5.1065	0.8423				
Cutoff 4 ("Agree"-"Strongly agree")	1.3321	0.5190				

Table B.6: Year 1 and Year 2 Samples Comparison - Selected Descriptive Statistics

	Year 1		Year 2	
	Mean	Std.Dev	Mean	Std.Dev
Obs.	2352		4452	
Student knowledge:				
3^{rd} grade math state test score	507.675	95.726	499.443	95.514
BAM test score (% correct)	53.453	22.123	53.787	21.633
Student characteristics:				
Age	9.52	0.50	8.91	0.81
Male	0.48		0.50	
White	0.26		0.23	
Black	0.43		0.46	
Hispanic	0.25		0.24	
Gifted	0.06		0.06	
Special education (SpEd)	0.09		0.11	
English language learner (ELL)	0.17		0.14	
Reduced price/free lunch	0.45		0.49	
Teacher characteristics:				
Obs.	177			
Years of experience in the district	6.40	5.94	5.42	4.59
Master's degree	0.53		0.46	
Teaches both Math and ELA (generalist)	0.83		0.82	
Classroom Composition				
Class size	23.29	4.97	24.15	6.29
Students baseline knowledge (3^{rd} grade test score):				
% low-level (1^{st} tercile)	30.20	21.80	35.20	25.70
% mid-level (2^{nd} tercile)	33.20	13.20	32.90	13.40
% high-level (3^{rd} tercile)	36.60	26.80	31.90	24.80

Table B.7: Within-Sample and Out-of-Sample Model Fit of Student Effort Measures

Within-Sample Fit										
	Never		Mostly not		Sometimes		Mostly		Always	
	Data	Model	Data	Model	Data	Model	Data	Model	Data	Model
I have done my best quality work in this class	0.007	0.009	0.010	0.013	0.089	0.116	0.242	0.306	0.457	0.556
In this class, I stop trying when the work gets hard	0.488	0.593	0.119	0.151	0.103	0.133	0.046	0.055	0.048	0.068
In this class, I take it easy and do not try to do my best	0.427	0.522	0.096	0.122	0.090	0.115	0.077	0.083	0.119	0.159
	None		Some		Most		All		All plus extra	
	Data	Model	Data	Model	Data	Model	Data	Model	Data	Model
How much homework do you usually complete?	0.006	0.007	0.062	0.082	0.106	0.138	0.489	0.604	0.137	0.169
Out-of-Sample Validation (Year 2 Data Sample Fit)										
	Never		Mostly not		Sometimes		Mostly		Always	
	Data	Model	Data	Model	Data	Model	Data	Model	Data	Model
I have done my best quality work in this class	0.007	0.009	0.021	0.012	0.010	0.112	0.291	0.302	0.431	0.566
In this class, I stop trying when the work gets hard	0.611	0.613	0.145	0.146	0.115	0.128	0.062	0.054	0.067	0.059
In this class, I take it easy and do not try to do my best	0.482	0.543	0.131	0.123	0.128	0.106	0.106	0.081	0.153	0.147
	None		Some		Most		All		All plus extra	
	Data	Model	Data	Model	Data	Model	Data	Model	Data	Model
How much homework do you usually complete?	0.008	0.007	0.062	0.078	0.147	0.135	0.554	0.603	0.229	0.177

Table B.8: Year 2 Sample - Student and Teacher Characteristics

	Year 1		Year 2			Year 1	Year 2
	Mean	Std.Dev	Mean	Std.Dev		Mean	Mean
Panel A: Students							
Obs.	2352		4452				
Age	9.52	0.50	8.91	0.81	Gifted	0.06	0.06
Male	0.48		0.50		Special education (SpEd)	0.09	0.11
White	0.26		0.23		English language learner (ELL)	0.17	0.14
Black	0.43		0.46		Reduced price/free lunch	0.45	0.49
Hispanic	0.25		0.24				
N. books in bedroom:					N. computers at home:		
None	0.09		0.08		None	0.12	0.11
≥1 and ≤10	0.22		0.21		One	0.45	0.41
≥11 and ≤24	0.21		0.21		More than one	0.43	0.48
≥25	0.48		0.50				
Has person at home to help with homework:					Has no quiet place to study at home:		
Never	0.02		0.03		Never	0.40	0.36
Mostly not	0.03		0.06		Mostly not	0.12	0.12
Sometimes	0.09		0.13		Sometimes	0.16	0.16
Mostly	0.17		0.14		Mostly	0.12	0.11
Always	0.69		0.64		Always	0.21	0.25
Panel B: Teachers							
Obs.	177						
Years of experience in the district	6.40	5.94	5.42	4.59			
Master's degree	0.53		0.46				
Teaches both Math and ELA (generalist)	0.83		0.82				
Classroom Composition							
Class size	23.29	4.97	24.15	6.29			
Student composition across classrooms:							
% low-level (1^{st} tercile)	30.20	21.80	35.20	25.70			
% mid-level (2^{nd} tercile)	33.20	13.20	32.90	13.40			
% high-level (3^{rd} tercile)	36.60	26.80	31.90	24.80			

Table B.9: Year 1 and Year 2 Samples - Teacher-Level Latent Factors Measures

	Year 1		Year 2	
	Mean	Std.Dev	Mean	Std.Dev
Teacher effort				
Teacher explains in another way if we do not understand (survey 0-4 score)	3.325	0.285	3.328	0.299
Teacher pushes us to work hard (survey 0-4 score)	3.092	0.370	3.192	0.427
Teacher does not waste time in class (survey 0-4 score)	2.664	0.385	2.709	0.402
Teacher asks us if we understand the lesson (survey 0-4 score)	3.329	0.315	3.392	0.322
Teacher asks us if we are following along (survey 0-4 score)	3.440	0.277	3.502	0.252
Teacher writes feedback on our papers (survey 0-4 score)	2.887	0.387	2.959	0.468
Teacher takes the time to summarize the lesson (survey 0-4 score)	2.813	0.480	2.981	0.434
Teacher encourage us to do our best (survey 0-4 score)	3.533	0.257	3.600	0.289
Teacher ability				
CLASS Behavior management scale	5.943	0.715	5.803	0.512
CLASS Content understanding scale	4.137	0.481	4.120	0.496
CLASS Productivity scale	5.918	0.555	5.803	0.419
FFTM Management of class procedures score	2.763	0.354	2.691	0.346
FFTM Management of student behavior score	2.840	0.344	2.767	0.380
MQI Richness of mathematics score	1.340	0.261	1.353	0.263
MQI Mathematical knowledge for teaching (MKT) score	2.030	0.218	2.027	0.225
Teacher explains clearly (survey 0-4 score)	3.321	0.295	3.324	0.269
Teacher controls class behavior (survey 0-4 score)	2.251	0.437	2.211	0.506
Teacher explains in orderly way (survey 0-4 score)	3.180	0.300	3.229	0.347
Teacher can explain in several ways (survey 0-4 score)	3.216	0.295	3.311	0.293
Teacher preference for adherence to standards				
Administrators require rigid adherence to standards	2.935	0.865	2.865	0.870
I frequently refer to and use information found in standards documents	2.355	0.592	2.341	0.631

Table B.10: Year 1 and Year 2 Samples - Student-Level Latent Factors Measures

	Year 1		Year 2	
Student knowledge:	Mean	Std.Dev	Mean	Std.Dev
3^{rd} grade math state test score	507.675	95.726	499.443	95.514
BAM test score (% correct)	53.453	22.123	53.787	21.633
Student effort				
I have done my best quality work in this class				
Never	0.007		0.007	
Mostly not	0.010		0.021	
Sometimes	0.089		0.104	
Mostly	0.242		0.291	
Always	0.457		0.577	
In this class, I stop trying when the work gets hard				
Never	0.488		0.611	
Mostly not	0.119		0.145	
Sometimes	0.103		0.115	
Mostly	0.046		0.062	
Always	0.048		0.067	
In this class, I take it easy and do not try to do my best				
Never	0.427		0.482	
Mostly not	0.096		0.131	
Sometimes	0.090		0.128	
Mostly	0.066		0.106	
Always	0.119		0.153	
How much homework do you usually complete?				
None	0.006		0.008	
Some	0.062		0.051	
Most	0.106		0.147	
All	0.489		0.554	
All plus extra	0.137		0.229	
Student preference for own knowledge				
I read at home almost every day				
Never	0.055		-	
Mostly not	0.079		-	
Sometimes	0.225		-	
Mostly	0.237		-	
Always	0.404		-	
School work is interesting				
Never	0.061		0.048	
Mostly not	0.072		0.061	
Sometimes	0.272		0.252	
Mostly	0.255		0.291	
Always	0.339		0.348	
School work is not very enjoyable				
Never	0.310		0.316	
Mostly not	0.159		0.171	
Sometimes	0.250		0.255	
Mostly	0.124		0.115	
Always	0.157		0.142	

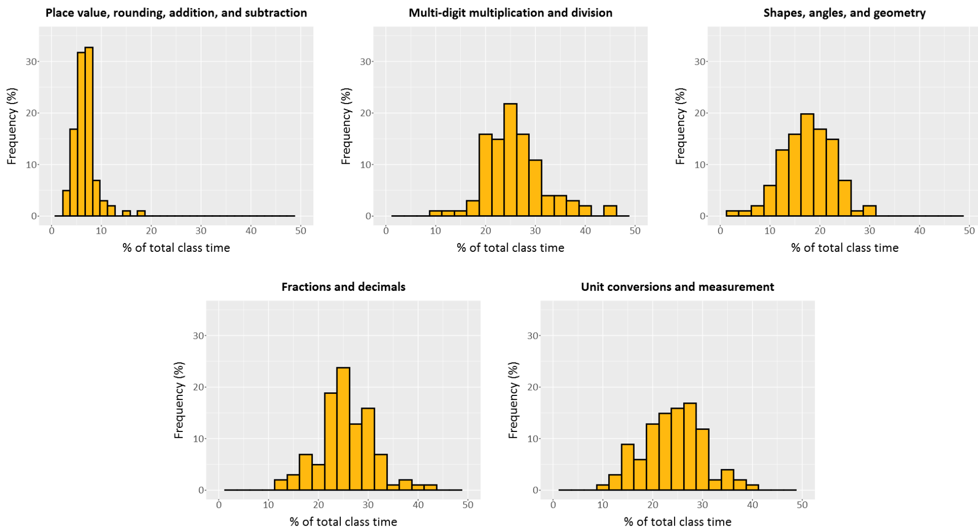

Figure B.1: Distribution of Class Time Allocations

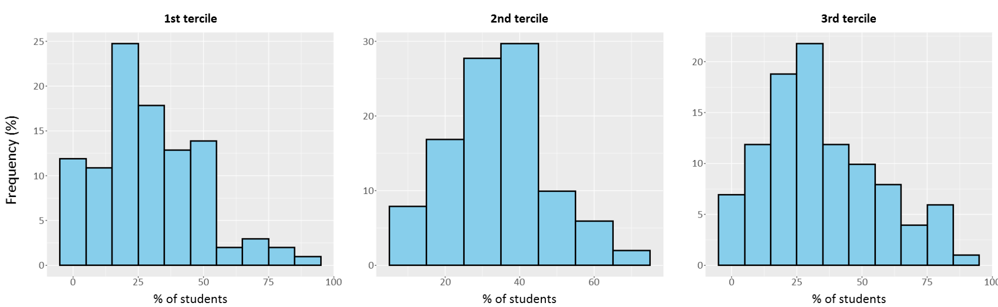

Figure B.2: Distribution of Classroom Composition

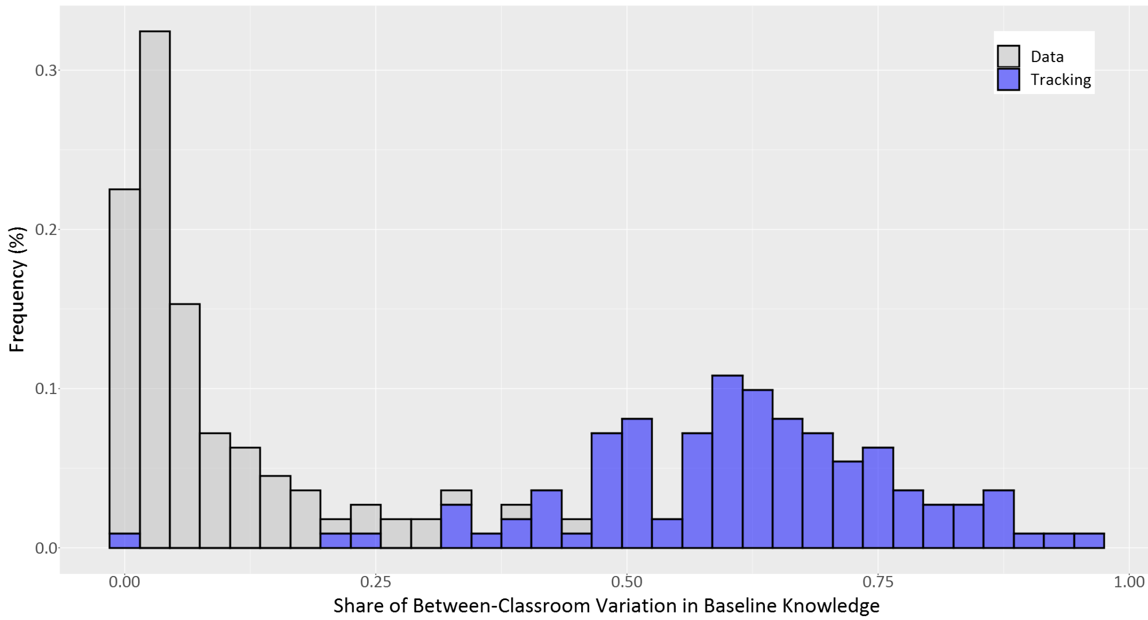

Figure B.3: Tracking Intensity Across Schools

Time on Topic		*Grades K-12 Mathematics Topics*	*Expectations for Students in Mathematics*				
<none>	1	**Number Sense/Properties/Relationships**	**Memorize Facts/ Definitions/ Formulas**	**Perform Procedures**	**Demonstrate Understanding of Mathematical Ideas**	**Conjecture/ Generalize/ Prove**	**Solve Non-Routine Problems/Make Connections**
⓪ ① ② ③	101	Place value	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	102	Whole numbers and integers	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	103	Operations	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	104	Fractions	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	105	Decimals	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	106	Percents	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	107	Ratios and proportions	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	108	Patterns	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	109	Real and/or rational numbers	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	110	Exponents and scientific notation	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	111	Factors, multiples, and divisibility	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③

Figure B.4: A snippet of the Survey of Enacted Curriculum for mathematics.