Accounting for Mathematics Performance of High School Students in Mexico: Estimating a Coordination Game in the Classroom

Petra Todd

University of Pennsylvania

Kenneth I. Wolpin

Rice University and University of Pennsylvania (Emeritus)

This paper estimates a model of the effort decisions of students and teachers in a classroom setting to understand the performance of Mexican high school students on curriculum-based examinations. The model allows for student heterogeneity in initial mathematics preparation and knowledge preference and for teacher heterogeneity in instructional ability and preferences for student knowledge. Survey data include multiple measurements of student and teacher effort, student and teacher preferences, student initial knowledge, and teacher ability. The most important factor accounting for poor performance, the lack of sufficient prior preparation, suggests a mismatch between the curriculum content and entering grade-level mathematics knowledge.

I. Introduction

There is concern in many countries that students are underachieving. This is particularly true in Mexico, where, on the basis of both international and national measures, the performance of students is poor. For

We thank Peter Arcidiacono, Flavio Cunha, Christopher Flinn, George Mailath, and Robert Miller for helpful discussions and five referees and the editor for comments. We have also benefited from workshop presentations at Bocconi University, Claremont-McKenna College, Collegio Carlo Alberto, Duke University, New York University, Harvard University, Queens

Electronically published October 29, 2018

[Journal of Political Economy, 2018, vol. 126, no. 6]

© 2018 by The University of Chicago. All rights reserved. 0022-3808/2018/12606-0010\$10.00

2609

example, Mexico ranked last among the 34 OECD countries in the 2009 PISA (Program for International Student Assessment) examination in mathematics.¹ Only 9.2 percent of ninth-grade students and 15.6 percent of twelfth-grade students scored at the proficient level or above on the 2008 ninth- and twelfth-grade national mathematics examination (ENLACE).²

There are many potential explanations for substandard test score performance related to student effort, teacher effort, teacher preparation (subject matter knowledge, teaching methods), school-level physical resources (libraries, textbooks, computers), the content of the curriculum, and the overall learning environment within the school (teacher morale, administrative leadership). Their relative importance has been debated, with initiatives focused on particular potential causes, for example, improving teacher training, redesigning curricula, and providing computers.

In this paper, we use control group data generated from a randomized controlled trial (the Aligning Learning Incentives [ALI] program) in Mexican federal high schools to assess the relative importance of some of these alternative explanations. In the ALI program, high school students in grades 10-12 were given end-of-year curriculum-based examinations in mathematics (see Behrman et al. [2015] for a detailed description of the ALI program). Students, teachers, and administrators in this treatment group were given monetary rewards of substantial magnitude depending on student performance on the ALI examination. The examinations were designed specifically for the ALI project by CENEVAL, a private nonprofit firm that is responsible for the design of a number of national tests. The tests are faithful to the content of the curriculum and test what students are supposed to know at the end of each school year. Table 1 reports the results of the examination in the final year of the program for tenth-grade students in the control group and in the treatment group for which the largest impact of the incentive on the test score was observed. Two conclusions emerge. First, the treatment effect in terms of the standardized score is large, 0.57 of a standard deviation. Second, neither the control group nor the treatment group mastered a significant part of the curriculum; the control group average raw score

University, the Stockholm School of Economics, University of California San Diego, University of Houston, University of Texas at Austin, University of Western Ontario, and Washington University. Funding from National Science Foundation grant SES-27364 is gratefully acknowledged. Data are provided as supplementary material online.

¹ PISA assessments, begun under the auspices of the OECD in 2000, are administered in reading, mathematics, and science to 15-year-olds. In 2009, 65 nations and territories participated.

² The percentage of students scoring at the proficient level (or above) on the ninthgrade test increased to 15.8 in 2011 and 20.6 in 2012 and on the twelfth-grade test to 24.7 and 30.8. The ENLACE, which had been used to provide teacher merit bonuses, has recently been abandoned in part because of questionable outcomes suggestive of improprieties in testing procedures.

TABLE 1
ALI TEST RESULTS

	Control Group	Treatment Group
Standardized score	500.5^{a}	556.8^{b}
Raw score	38.3	44.9°
Number of correct answers ^d	30.3	35.5

NOTE.—The results are for the tenth grade in the last (third) year of the ALI program. Estimated treatment effect uses the same specification as in Behrman et al. (2015) on a slightly different sample. The treatment group corresponds to T3 in Behrman et al.'s study, the group in which students, teachers, and administrators were compensated for student performance.

^a Standard deviation = 100.

 $^{\rm b}$ The standard error of the difference from the control group, 56.3, is 7.19.

^c The standard error of the difference from the control group, 6.6, is 0.83.

^d There were 79 multiple-choice questions with four possible answers.

was 38.3 percent correct and the treatment group average 44.9 percent correct.³ Following the same procedure as for the ENLACE, the experts consulted by CENEVAL placed test scores into four categories: prebasic, basic, proficient, and advanced. In that metric, 80 percent of the students in the control group had achieved only a prebasic level of tenth-grade mathematics knowledge, 20 percent had achieved a basic level, and none of the students had achieved a level considered proficient or advanced.⁴

The first-order question raised by these results is why these Mexican high school students master so little of the curriculum. This paper provides a framework within which to assess quantitatively the reasons. For that purpose, we develop and estimate a model of a classroom in which multiple students and a single teacher make effort decisions that affect student performance. Students begin the academic year with an initial knowledge level, which, in combination with their effort and their teacher's effort during the year as well as their teacher's instructional ability, produces end-of-year knowledge. Teacher effort, assumed to be a pure public input (consistent with a conventional lecture style of instruction as is normal practice), and student effort are complementary in producing student knowledge. Class size potentially affects the marginal products of student and teacher effort. Schools may differ in their production technologies. The value-added specification of the cognitive achievement production function we adopt is derived from a cumulative specification (see Todd and Wolpin 2003) in which current knowledge depends on all past inputs, on student ability, and on initial knowledge at

³ The ALI test consisted of 79 multiple-choice questions with four possible answers. Most of the questions were word problems, which generally would require more than mere memorization.

⁴ In the treatment group, 55 percent of the students were in the prebasic category and the rest in the basic category.

the start of formal schooling. Given the cumulative specification, initial knowledge at the start of any grade is a sufficient statistic for this history.

Students have preferences over end-of-year knowledge, while teachers care about the sum of the end-of-year knowledge levels of their students. Student and teacher effort are assumed to be chosen within a Nash game in which each maximizes utility net of a variable effort cost. Given production complementarities between student and teacher effort, all students in the class benefit from an increase in any one student's effort through the induced increase in teacher effort. As in coordination games more generally, there are multiple equilibria.⁵ In particular, the model exhibits two equilibria, one in which all students and the teacher exert positive effort (the positive-effort equilibrium) and one in which all students and the teacher exert zero effort (the zero-effort equilibrium).⁶ Given our specification of the production function and student and teacher preferences, student and teacher effort have closed-form solutions in the positive-effort equilibrium. The solution to the game is assumed to be unique in the all-positive equilibrium, because the all-zero equilibrium is not consistent with the effort data.

The ALI project conducted surveys of students and teachers to collect data that would help in understanding the mechanisms governing student performance, in particular, data on student and teacher effort, student and teacher preferences, teacher instructional ability, and student mathematics knowledge at the start of the school year. We develop a simulated maximum likelihood procedure that uses multiple measures of effort and multiple measures and exogenous determinants of student and teacher endowments, that is, student beginning-year mathematics knowledge, teacher instructional ability, and student and teacher preferences for mathematics knowledge.⁷ In addition, the estimation incorporates unobserved heterogeneity in productivity across schools.

The estimation recovers the parameters of the knowledge production function, the utility functions of students and teachers, and the measurement system. In principle, the measurement system provides sufficient structure to recover the parameters of the knowledge production function without solving the effort game. However, as long as student and teacher effort are subject to choice, estimates of the production function

⁷ This econometric framework has antecedents in the MIMIC (multiple-indicator multiple-cause) framework (see, e.g., Joreskog and Goldberger 1975). For recent applications and extensions, see Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010).

⁵ See Vives (2005) for a discussion of games with strategic complementarities.

⁶ The positive-effort equilibrium is obtained as the solution to the reaction functions. The solution contains students who optimally choose zero effort because they have a zero preference for mathematics knowledge. The positive-effort equilibrium is defined to include those students.

alone cannot determine the underlying factors responsible for low student performance. Estimates of the entire structure, the production function, and the utility function, obtained by solving the effort game for the effort functions, are necessary to distinguish among the potential causes.

There are additional reasons to estimate the effort functions along with the production function. Because the production function and the effort functions depend on the same parameters, the model imposes cross-equation restrictions. We take advantage of the overidentification that arises from these restrictions in two ways. First, we identify additional parameters in the measurement structure; in particular, we estimate parameters that reflect the amount of reported effort of students and teachers that is unproductive (i.e., has zero marginal product). We are thus able to determine the extent to which "true" effort diverges from reported effort, either because of biased reporting or because effort is actually productive only above some threshold. Second, we test overidentifying restrictions of the model.⁸

On the surface, it would not appear that a major cause of poor performance is a lack of student or teacher effort. On average, students report spending 4.7 hours per week studying math outside of class and teachers report spending 4.0 hours per week preparing lessons. However, we find that "true" effort is considerably smaller, on average, only 3.1 hours per week (about 45 minutes per class hour) for students and 1.6 hours per week (about 25 minutes per class hour) for teachers. Despite these seemingly low levels of effort, our results indicate that insufficient student or teacher effort does not significantly contribute to the lack of mastery of the tenth-grade mathematics curriculum. Nor do we find low teacher ability to be an important contributing factor.

The factor that emerges from the empirical analysis to be the most important in determining a student's tenth-grade mathematics knowledge is his or her previous mathematics knowledge in ninth grade. On the basis of our estimates, we simulate that if the ninth-grade mathematics knowledge of all students were two standard deviations (SD) above the mean, then the tenth-year mathematics knowledge would be, on average, 1.7 SD above the mean. Although this is a large effect in relative terms, the average raw test score, even in this case, would remain below 50 percent. Our production function estimates show that the marginal product of student effort is very small. An increase in student effort from

⁸ Cross-equation restrictions arise for the same reason as in the joint estimation of firm production functions and input demand equations, namely, that the same parameters appear in both. The analogy is not exact because input demand, i.e., student and teacher effort, is determined also by preference parameters. However, we have multiple measures of student and teacher preferences that pin down those parameters.

zero effort to effort that is 2 SD above the mean (an increase of about 5 hours per week in study time) increases tenth-grade mathematics knowledge by only 0.10 SD for students whose performance in ninth-grade mathematics is at the mean. We find similar size effects for the marginal product of teacher effort.

A number of possible explanations are consistent with these results. The most convincing, in our view, is that the tenth-grade mathematics curriculum is too difficult given the mathematics knowledge that students have acquired through the ninth grade. Moreover, the finding that student effort has only a small effect on tenth-grade knowledge even for students who had the highest mathematics performance in the ninth grade, as measured by their ENLACE score and the grade they received in their ninth-grade mathematics course, would seem to imply that those students did not actually master the ninth-grade curriculum. Presumably, mastery of the tenth-grade curriculum. Although speculative, it would thus seem that the ninth-grade curriculum is also too difficult for students to acquire the requisite knowledge to be successful in tenth grade.⁹

There are only a few previous studies that develop explicit models of teacher or student effort choices and, to our knowledge, none that implement a model of both student and teacher effort choices. Duflo, Dupas, and Kremer (2008) develop a model in which teachers choose effort levels and a target level at which to orient their instruction, taking into account their students' previous performance levels. They test the model's implications using data from a tracking experiment in Kenya that randomly assigned some schools to a treatment in which classroom assignment depended on prior performance. They find that teacher effort, measured by teacher attendance, is higher under the tracking regime and that both high- and low-ability students benefit from tracking in terms of performance. In another study, Duflo, Hanna, and Ryan (2012) develop a dynamic model of teacher attendance in India to study compensation schemes and implications for student performance using data from a randomized experiment.¹⁰ They find that financial incentives increase teacher attendance and improve student test scores. These models do not, however, incorporate student effort. Kremer, Miguel, and Thornton (2009) reference, but do not explicitly develop, a model of strategic effort complementarities to interpret the results of a ran-

⁹ This explanation is consistent with that given by Pritchett and Beatty (2012), who argue that the slow pace of learning evident in a number of developing countries is the result of reliance on overly ambitious curricula.

¹⁰ This paper builds on a growing literature that combines structural estimation with randomized controlled experiments (see Todd and Wolpin [2006] and the citations within).

domized merit scholarship program in which Kenyan girls received school fees and a grant depending on academic exam performance. Although the incentives were provided only to high-performing girls, they find that girls with low pretest scores, who were unlikely to win the tournament, also showed improvement in performance as did boys. Also, teachers in the schools assigned to the scholarship program had higher attendance. They note that these results are consistent with there being positive classroom externalities to study effort and potentially a strategic complementarity between student and teacher effort. The model developed and estimated in this paper incorporates this kind of complementarity and can potentially be helpful in explaining the pattern of experimental results in their study.

A related literature develops models in which peer group norms influence individuals' educational investment choices (e.g., Brock and Durlauf 2001; Austen-Smith and Fryer 2005). Lazear (2001) considers a different type of model in which one disruptive student imposes negative spillovers on other students' learning in a classroom, and he uses the model to study implications for optimal class size. In our model, there are also spillover effects of peers on other students' learning, but these spillover effects arise indirectly through teacher effort choices.¹¹

This paper also is related to the empirical literature on the estimation of models with strategic complementarities. Examples include models of the adoption by banks of the automated clearinghouse system in Ackerberg and Gowrisankaran (2006), the timing of desertions during the Civil War in De Paula (2009), and the timing of radio commercials in Sweeting (2009). In contrast to these applications, in the model we estimate the objects of choice (student and teacher effort levels) are continuous rather than discrete and we assume complete information.

The paper proceeds as follows. Section II presents the model, Section III describes the estimation procedure, and Section IV discusses the data. Section V presents the results and a discussion of their interpretation. Section VI presents a summary.

II. The Model: Structure

This section presents a model of the production of student knowledge within a classroom setting. End-of-year student knowledge depends on the student's initial level of knowledge and effort and on the teacher's

¹¹ A more general formulation might also incorporate direct peer-to-peer spillover effects as well as spillover effects arising through teacher effort choices. Such a specification would raise an identification issue and substantially complicate the solution of the coordination game. We therefore leave development and estimation of such a model for future work.

instructional ability and effort. Student and teacher effort levels are assumed to be the outcome of a Nash game.¹²

Consider a class j in grade g with a single teacher t and N_j students. Each student, n, begins with an initial level of knowledge, $K_{g-1,nj}$, $n = 1, ..., N_j$, and chooses a level of learning effort, ε_{gnj} . The teacher has instructional ability a_{gt} and chooses instructional effort ε_{gij} , a pure public input (the same for each student). End-of-year knowledge in grade g for student n, K_{gn} (dropping the j subscript), is produced according to

$$K_{gn} = \delta_0 K_{g-1,n} \cdot \left(1 + \kappa K_{g-1,n}^{\delta_1} a_{gt}^{\gamma_0} \varepsilon_{gn}^{\gamma_1} \varepsilon_{gt}^{\gamma_2} \right), \tag{1}$$

where $\delta_0 K_{g-1,n} \ge 0$ is the level of knowledge achieved if either the student or teacher chooses zero effort and κ is a normalization that converts units.¹³ The Cobb-Douglas component of (1) represents the proportionate increase in knowledge due to student and teacher effort over the level of knowledge that remains if the student or teacher exerts zero effort.

In the value-added specification given in (1), $K_{g-1,n}$, knowledge in the previous grade, is taken to be a sufficient statistic for student ability, for student and teacher effort in all prior grades, and for parental preschool home inputs. A feature of our specification is that it can be derived from a cumulative specification in which current knowledge depends on all past inputs and student ability.¹⁴ In particular, this cumulative specification of the production function for knowledge produced in grade g is

$$K_{gn} = \tau_g f(A_n, K_{0n}) \prod_{s=1}^g \left(1 + \kappa K_{s-1,n}^{\delta_1} a_{st}^{\gamma_0} \varepsilon_{sn}^{\gamma_1} \varepsilon_{st}^{\gamma_2} \right), \tag{2}$$

where A_n is the student's ability, K_{0n} the student's preschool knowledge (itself a function of preschool inputs), and τ_g one minus a depreciation rate. Note that $\tau_g f(A_n, K_{0n})$ is the level of knowledge a student would have at the end of grade g if the student (or teachers) were to exert zero effort in every grade. Dividing K_{gn} by $K_{g-1,n}$ leads to the value-added specifica-

¹³ In order to capture differences in unobserved factors affecting student outcomes at the school level, we allow schools to differ in δ .

¹⁴ Value-added specifications are often adopted in the literature to circumvent the need for historical data on inputs and ability.

¹² An alternative assumption is that the teacher is a Stackelberg leader. Nash is chosen in part because we have data only about student and teacher effort averages over a school year. Students and teachers interact on a daily basis in a repeated situation in which students and teachers learn over time about each other's characteristics and effort. Allowing teachers to react to student effort (Nash) would seem a better approximation to that situation than assuming that teachers make a one-time decision for the entire school year (Stackelberg). Indeed, Bagwell (1995, 272) has shown that "even with only the slightest degree of imperfection in the observability of the first-mover's selection . . . the pure-strategy Nash equilibrium outcomes for the associated simultaneous-move game."

tion given in (1) with $\delta_0 = \tau_g / \tau_{g-1}$.¹⁵ In what follows, we drop the *g* subscript, because our analysis is conducted on a single grade.

Students are assumed to maximize their utility from knowledge net of effort cost (c_n) :

$$U_n(\varepsilon_n) = \theta_n K_n - \frac{c_n}{2} (\varepsilon_n)^2, \qquad (3)$$

where θ_n is student *n*'s (constant) marginal utility of knowledge. The teacher is assumed to care about the total amount of knowledge produced in the class. Given effort cost, c_i , the teacher maximizes

$$U_t(\varepsilon_t) = \phi_t \sum_{n=1}^N K_n - \frac{c_t}{2} (\varepsilon_t)^2, \qquad (4)$$

where ϕ_i is teacher *t*'s (constant) marginal utility of total student knowledge.

Assuming that all student endowments, $K_{-1,n}$, θ_n , and c_n for all n = 1, ..., N, and teacher endowments, $a_b \phi_b$ and c_b are public information, the reaction functions for the Nash equilibrium game are

$$\varepsilon_n = \left(\gamma_1 \kappa \delta_0 K_{-1,n}^{1+\delta_1} a_t^{\gamma_0} \theta_n c_n^{-1}\right)^{\frac{1}{2-\gamma_1}} \varepsilon_t^{\frac{\gamma_2}{2-\gamma_1}},\tag{5}$$

$$\varepsilon_t = \left(\gamma_2 \kappa a_t^{\gamma_0} \phi_t c_t^{-1}\right)^{\frac{1}{2-\gamma_2}} \left(\sum_{n=1}^N \delta_0 K_{-1,n}^{1+\delta_1} \varepsilon_n^{\gamma_1}\right)^{\frac{1}{2-\gamma_2}} \tag{6}$$

for n = 1, ..., N.

The unique solution to (5) and (6), normalizing c_n and c_i to one, has a closed form given by

$$\varepsilon_{n}^{*} = \gamma_{1}^{\frac{2-\gamma_{2}}{4-2(\gamma_{1}+\gamma_{2})}} \gamma_{2}^{\frac{\gamma_{2}}{4-2(\gamma_{1}+\gamma_{2})}} \kappa_{t}^{\frac{2}{4-2(\gamma_{1}+\gamma_{2})}} a_{t}^{\frac{2\gamma_{0}}{4-2(\gamma_{1}+\gamma_{2})}} \phi_{t}^{\frac{\gamma_{2}}{4-2(\gamma_{1}+\gamma_{2})}} (\delta_{0}K_{-1,n})^{\frac{1}{2-\gamma_{1}}} \times \left[\sum_{n=1}^{N} (\delta_{0}K_{-1,n})^{\frac{2}{2-\gamma_{1}}} K_{-1,n}^{\frac{2\delta_{1}}{2-\gamma_{1}}} \theta_{n}^{\frac{\gamma_{1}}{2-\gamma_{1}}} \right]^{\frac{\gamma_{2}}{4-2(\gamma_{1}+\gamma_{2})}},$$
(7)

¹⁵ The specification of the value-added production function is also consistent with the conventional Ben-Porath (1967) production function in which schooling is a period of specialization in producing human capital. The Ben-Porath formulation is

$$\dot{K}_g = \beta_0 (s_g K_g)^{\beta_1} D_g^{\beta_2} - \delta K_g,$$

where \dot{K}_g is the time (grade) derivative of knowledge, s_g is the fraction of time devoted to investing in knowledge, D_g are other inputs, and $1 - \delta$ is the depreciation rate. Setting $s_g = 1$ (specialization), $\beta_1 = \delta_1$, $D_g = \kappa a_{dg}^{\gamma_0} \varepsilon_{dg}^{\gamma_1} \varepsilon_{dg}^{\gamma_2}$, and dividing through by K_g , one gets the equivalent of (1) in the text, where $\beta_0 = \delta_0 \cdot \kappa$.

$$\varepsilon_{t}^{*} = \gamma_{1}^{\frac{2-\gamma_{1}}{4-2(\gamma_{1}+\gamma_{2})}} \gamma_{2}^{\frac{2-\gamma_{1}}{2}(\gamma_{1}+\gamma_{2})} \kappa^{\frac{2}{4-2(\gamma_{1}+\gamma_{2})}} a_{t}^{\frac{2\gamma_{10}}{4-2(\gamma_{1}+\gamma_{2})}} \phi_{t}^{\frac{2-\gamma_{1}}{4-2(\gamma_{1}+\gamma_{2})}} \\ \times \left[\sum_{n=1}^{N} (\delta_{0} K_{-1,n})^{\frac{2}{2-\gamma_{1}}} K_{-1,n}^{\frac{2\delta_{1}}{2-\gamma_{1}}} \theta_{n}^{\frac{\gamma_{1}}{2-\gamma_{1}}} \right]^{\frac{2-\gamma_{1}}{4-2(\gamma_{1}+\gamma_{2})}}.$$
(8)

As seen, a student's effort depends not only on the student's own attributes but also, through the teacher's effort decision, on the attributes of all of the other students in the class. In this sense, there are student peer effects on end-of-year knowledge, although they operate through the teacher's effort decision. Given that the attributes of all of the students in the class are taken into account by the teacher in choosing effort, separately identifying the effect of peer characteristics in the knowledge production function (e.g., the initial knowledge of classmates) from the effect of teacher effort requires a functional form assumption about how peers enter the production function.¹⁶

An implication of the assumption that teachers care about the sum of student knowledge is that teacher effort (and student effort) is increasing in class size. Thus, students in larger classes will acquire more knowledge. As this implication may not be consistent with the data, we allow for the marginal products of student and teacher effort to be affected by class size; specifically, we set $\gamma_1 = \gamma_{10} + \gamma_{11}N$ and $\gamma_2 = \gamma_{20} + \gamma_{21}N$. Negative values for γ_{11} and γ_{21} would offset the implied positive impact of class size due to the assumption about teacher utility.

It is always the case, regardless of parameter values, that there is an equilibrium in which all students and the teacher exert zero effort. If the teacher exerts zero effort, it will not be optimal for any single student (or for all or any subset of students) to exert positive effort, given that the marginal product of student effort is zero. Similarly, if all students exert zero effort, it will not be optimal for the teacher to exert positive effort, given that the marginal product of teacher effort is zero. Thus, there are two equilibria. Clearly, the equilibrium in which all students and the teacher exert zero effort is Pareto inferior to the equilibrium in which the students and the teacher exert positive effort. We will argue below, on the basis of the data, that the zero equilibrium is never selected.

¹⁶ Fruehwirth (2014) discusses estimation of models with direct peer-to-peer effects. She shows that when peer effects arise from unobservable factors, then the types of specifications commonly estimated in the literature are of limited use to inform policies (e.g., class-room regrouping policies or desegregation). Our framework offers an alternative approach that would allow simulation of the effects of such policies, although it assumes that the effort decisions generating the peer effects can be measured with error.

III. Estimation

Heterogeneity among students within a class potentially arises from differences in initial knowledge and preferences, $K_{-1,n}$ and θ_n . Heterogeneity among students across classes in the same school and across schools arises from class- or schoolwide differences in $K_{-1,n}$ and θ_n . Heterogeneity among teachers within the same school or among schools arises from differences in ϕ_t and a_r . Heterogeneity in outcomes across schools also arises from differences in technology; specifically, the production function parameter, δ_0 , is assumed to be a school-specific random coefficient.

A. Latent Factor Structure for Student and Teacher Endowments

Student and teacher endowments, K_{0n} , θ_n , ϕ_n , and a_n , are treated as latent factors that are measured with error. To be concrete, assume there are h = 1, ..., H schools, with $j = 1, ..., J_h$ classes in school h and $n = 1, ..., N_{jh}$ students in class j in school h. We let beginning knowledge of student n assigned to class j in school h ($K_{-1,njh}$) depend on a set of exogenous initial conditions ($X_{njh}^{K_{-1}}$) and potentially on school-, class-, and individual-level error components:

$$K_{-1,njh} = X_{njh}^{K_{-1}} \beta^{K_{-1}} + \xi_h^{K_{-1}} + \mu_{jh}^{K_{-1}} + \omega_{njh}^{K_{-1}}.$$
 (9)

The first error component, $\xi_{h}^{K_{-1}}$, allows for unobserved school-level differences in student initial knowledge; the second, $\mu_{jh}^{K_{-1}}$, for unobserved classlevel differences within a school; and the third, $\omega_{ijh}^{K_{-1}}$, for idiosyncratic withinclass differences.¹⁷ Student preferences follow a similar structure:

$$\theta_{njh} = X^{\theta}_{njh}\beta^{\theta} + \xi^{\theta}_{h} + \mu^{\theta}_{jh} + \omega^{\theta}_{njh}.$$
 (10)

All error components in (9) and (10) are assumed to be mean zero and to be orthogonal to each other and to observed characteristics.

Assume there are $t = 1, ..., T_h$ teachers in school h, where $T_h \leq J_h$. If two or more classes have the same teacher, both ϕ_i and a_i will be the same in the two classes. Teacher instructional ability depends on observable teacher characteristics (X_{ih}^a) and unobservables at the school level (ξ_h^a) and at the teacher level (μ_{ih}^a) , namely,

$$a_{th} = X^a_{th}\beta^a + \xi^a_h + \mu^a_{th}.$$
 (11)

Similarly, teacher preferences are given by

$$\boldsymbol{\phi}_{th} = X^{\phi}_{th} \beta^{\phi} + \xi^{\phi}_{h} + \mu^{\phi}_{th}. \tag{12}$$

¹⁷ In estimation, given the number of latent variables, we do not allow for error components at all levels to conserve on parameters.

School- and teacher-level error components are assumed to be orthogonal to each other and to observable characteristics.¹⁸ Teacher and student error components can be freely correlated across endowments.

B. Measurement Structure

In terms of the latent endowments, there are assumed to be M^j measures for prior knowledge, student preferences, teacher ability, and teacher preferences ($j = K_{-1,n}$, θ_{njh} , a_{th} , ϕ_{th}). The measurement equations are given by

$$K_{-1,n}^{m} = \alpha_{0}^{K_{-1,m}} + \alpha_{1}^{K_{-1,m}} K_{-1,n} + \zeta_{njh}^{K_{-1,m}}, \quad m = 1, \dots, M^{K_{-1}},$$
(13)

$$\theta_{njh}^{m} = \alpha_{0}^{\theta,m} + \alpha_{1}^{\theta,m} \theta_{njh} + \varsigma_{njh}^{\theta,m} \quad \text{for } m = 1, \dots, M^{\theta_{n}},$$
(14)

$$\phi_{th}^{m} = \alpha_{0}^{\phi,m} + \alpha_{1}^{\phi,m} \theta_{th} + \varsigma_{th}^{\phi,m} \quad \text{for } m = 1, \dots, M^{\phi_{t}}, \tag{15}$$

$$a_{th}^{m} = \alpha_{0}^{a,m} + \alpha_{1}^{a,m} a_{th} + \zeta_{th}^{a,m} \quad \text{for } m = 1, \dots, M^{a_{t}}, \tag{16}$$

where the different measurements of each of the latent factors are denoted with an *m* superscript.

Student and teacher effort are the outcomes of the effort game, which are fully determined by the latent endowments. We also treat them as latent variables that are measured with error.¹⁹ There are assumed to be M^{ε_1} measures of student effort and M^{ε_2} measures of teacher effort. The effort measurement equations take the form

$$\varepsilon_{njh}^{m} = \alpha_{0}^{\varepsilon_{1},m} + \alpha_{1}^{\varepsilon_{1},m} \varepsilon_{njh} + \varsigma_{njh}^{\varepsilon_{1},m} \quad \text{for } m = 1, \dots, M^{\varepsilon_{1}}, \tag{17}$$

$$\varepsilon_{ijh}^{m} = \alpha_{0}^{\varepsilon_{2},m} + \alpha_{1}^{\varepsilon_{2},m} \varepsilon_{ijh} + \varsigma_{ijh}^{\varepsilon_{2},m} \quad \text{for } m = 1, \dots, M^{\varepsilon_{2}}.$$
(18)

There is one measure of end-of-year knowledge, a test score T_{njh} , which we assume measures knowledge with an additive measurement error

$$T_{njh} = K_{njh} + \zeta_{njh}^{T}, \tag{19}$$

where K_{njh} is determined by (1). All measurement errors, for the endowments, for student and teacher effort, and for end-of-year knowledge,

¹⁸ All of the student and teacher latent factors are bounded from below by zero. Thus, we assume that no student nor any teacher receives negative utility from knowledge.

¹⁹ This methodology allows us to reduce the dimension of the effective input space to one for both the student and the teacher. See Cunha and Heckman (2008) for further discussion.

are assumed to be uncorrelated with all of the latent variables (both observable and unobservable components) and with each other.

An observation consists of (i) measures of the effort levels, end-of-year test scores, and measures of initial knowledge and preferences for each of the N_{jh} students in each class *j* and school *h*; and (ii) measures of the effort level, preference, and ability of the teacher in each class. We denote the observation set for class *j* (and teacher *t*) in school *h* as

$$O_{jh} = \left\{ \varepsilon^m_{njh}, \varepsilon^m_{tjh}, T_{njh}, K^m_{-1,njh}, \theta^m_{njh}, \phi^m_{th}, a^m_{th} : n = 1, \dots, N_{jh} \right\}$$

for all measures (i.e., over all *m*) and all classrooms in all schools.

In describing the distributional assumptions, it is convenient to collect terms. Let $\Omega_1 = \{X_{njh}^{K_{-1}}, X_{njh}^{\theta}, X_{ih}^{a}, X_{ih}^{\phi}\}$ denote the vector of observable characteristics of the students and teacher in class *j* of school *h*, $\Omega_2 = \{\xi_h^{K_{-1}}, \xi_h^{\theta}, \xi_h^{a}, \xi_h^{\phi}, \delta_{0h}\}$ the vector of school-level unobservables, $\Omega_3 = \{\mu_{jh}^{K_{-1}}, \mu_{jh}^{\theta}, \mu_{ih}^{a}, \mu_{ih}^{\theta}\}$ the vector of within-school class- and teacher-level unobservables, $\Omega_4 = \{\omega_{njh}^{K_{-1}}, \omega_{njh}^{\theta}\}$ the vector of within-class student-level unobservables, and $\Omega_5 = \{\varsigma_{njh}^{K_{-1}}, \varsigma_{njh}^{\eta}, \varsigma_{njh}^{\psi_{in}}, \varsigma_{njh}^{\psi_{in}}, \varsigma_{njh}^{\phi_{in}}, \varsigma_{nh}^{\phi_{in}}\}$ the vector of measurement errors. The unobservables in Ω_i (*i* = 2, 3, 4) each have joint distribution F_i , assumed to be normal with variance-covariance matrix Λ_i . The joint distribution is tribution of the measurement errors, denoted as F_5 , is also assumed to be normal with (diagonal) variance-covariance matrix Λ_5 .

C. Likelihood Function

Estimation is carried out by simulated maximum likelihood. The likelihood contribution for students $n = 1, ..., N_{jh}$ in class j of school h is the joint density of O_{jh} , that is, the measured efforts of students and teachers, students' end-of-year (ALI) test scores, measured student preferences, and measured teacher abilities and preferences. For now, we ignore the normalizations that are necessary for identification.

The estimation procedure is as follows:

1. Choose a set of parameter values

$$\{\kappa,\gamma_0,\gamma_1,\gamma_2,ar{\delta}_0,\sigma^2_{\delta_0},eta^{K_{-1}},eta^ heta,eta^ heta,eta^a,\Lambda_2,\Lambda_3,\Lambda_4,\Lambda_5,\Gamma\},$$

where Γ denotes the α_0 and α_1 parameters in the measurement error equations.

2. Draw school-level shocks, Ω_2 , for each school, h = 1, ..., H, classand teacher-level shocks, Ω_3 , for each class and teacher in all schools, and student-level shocks, Ω_4 , for each student in all classes and schools.

3. Given the shocks drawn in (2) and the set of observable variables (Ω_1) , calculate each student's value of $K_{-1,njh}$ and θ_{njh} and each teacher's value of a_{th} and ϕ_{th} (setting to zero any latent factor that is negative).

4. Calculate the equilibrium from (7) and (8) for each class and for all schools for each of the d = 1, ..., D draws. Each equilibrium is characterized by the optimal student and teacher effort levels and implied end-of-year knowledge.

5. For each draw (*d*), and given the joint measurement error distribution, calculate the joint likelihood of observing all of the measured variables for the students and the teacher, which is given by the joint density of measurement errors.²⁰ Denote this product for classroom *j* in school *h* as $\hat{f}_{ijh}(d)$ for d = 1, ..., D.²¹

The likelihood for class j in school h is the average over the likelihoods for each draw, namely,

$$L_{jh} = \frac{1}{D} \sum_{d} \hat{f}_{jh}(d).$$
(20)

7. Repeat for all J_h classes in school h and over all h = 1, ..., H schools. The likelihood over the entire sample is

$$\prod_{h=1}^{H} \prod_{j=1}^{J_h} L_{jh}.$$
(21)

8. Repeat 1–7, maximizing (21) over the parameter vector given in step 1.

D. Identification

We discuss identification separately for different sets of parameters.

i. The parameters in the distribution of latent endowments, K_{-1} , θ , a, ϕ , and the measurement error distribution: Multiple measurements of the latent endowments together with observable determinants, (13) and (9) for student initial knowledge, (14) and (10) for student preferences, (16) and (11) for teacher instructional ability, and (15) and (12) for teacher

²⁰ Some of the measures are continuous, some are ordered categorical, and some are dichotomous, and all of the continuous measures are bounded. In all cases, we assume there is a continuous latent measure that underlies the observed measure. Thus, bounded measures are treated as truncated, dichotomous variables as probits, and ordered categorical variables as ordered probits.

²¹ Because all measurement equations are linear, the Jacobian of the transformation from the observed measures to the measurement errors is one.

preferences, identify the joint distribution of the latent endowments, with suitable normalizations, and the measurement error distributions.²²

ii. Production function parameters: $\bar{\delta}_0$, $\sigma_{\delta_0}^2$, δ_1 , κ , γ_0 , γ_1 , γ_2 . Combining (19) and (1), the parameters of the production function are estimated from

$$T_n = \delta_{0h} K_{-1,n} \cdot \left(1 + \kappa K_{-1,n}^{\delta_1} a_t^{\gamma_0} \varepsilon_n^{\gamma_1} \varepsilon_t^{\gamma_2} \right) + \varsigma_n^T.$$

$$\tag{22}$$

To fix ideas, suppose we have perfect measures of $K_{-1,n}$, a_i , ε_n , and ε_i . Given that the only student-level unobservable is the test score measurement error, which is assumed to be orthogonal to the determinants of K_n , identification of the production function parameters is immediate given independent variation in the observables. Within-school variation (where there are multiple teachers) identifies all of the parameters except for $\sigma_{\delta_0}^2$, which is identified from between-school variation in test scores.

We do not assume to have perfect measures of $K_{-1,n}$, $a_{\nu} \varepsilon_{\nu}$ and ε_{n} . However, with multiple measures, as we have, we can apply a theorem in Cunha et al. (2010) to demonstrate identification. In fact, our formulation is a special case of the theorem, and we can apply the theorem to simultaneously identify the distribution of latent endowments (as in point i above), together with the production function parameters. Specifically, define the vector of the determinants of K_n (end-of-year knowledge), $K_{-1,n}$, ε_n , a_{ν} , ε_t in (22) as χ_n . We can treat the test score as a separate measure of the vector, where the measurement is given by the nonlinear function

$$T_n = K_n(\boldsymbol{\chi}_n; \boldsymbol{\delta}_0, \sigma_{\boldsymbol{\delta}_0}^2, \boldsymbol{\delta}_1, \boldsymbol{\kappa}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) + \boldsymbol{\varsigma}_n^T.$$
(23)

 $^{\ensuremath{\scriptscriptstyle 22}}$ For example, consider two measures of student initial knowledge, with measurement equations

$$\begin{split} K_{-1,n}^{m1} &= K_{-1,n} + \varsigma_n^{K_{-1,n}}, \\ K_{-1,n}^{m2} &= \alpha_0^{K_{-1,n2}} + \alpha_1^{K_{-1,n2}} K_{-1,n} + \varsigma_n^{K_{-1,n2}}, \end{split}$$

where, for convenience, we allow only for a student-level unobservable. Note that we have normalized $\alpha_0^{K_{-1,ni}} = 0$ and $\alpha_1^{K_{-1,ni}} = 1$, which establishes the first measure as the metric for $K_{-1,n}$. Initial knowledge is given by

$$K_{-1,n} = \beta_0^{K_{-1}} + \beta_1^{K_{-1}} X_n^{K_{-1}} + \omega_n^{K_{-1}}$$

where, to simplify, $X_n^{K_1}$ is a scalar and there is only a student-level unobservable. All unobservables are orthogonal to each other and to X. Clearly $\beta_1^{K_1}$ and $\beta_1^{K_1}$ are identified from the first measurement equation after substitution for $K_{-1,n}$ above, that is, $\beta_1^{K_1}$ by the regression coefficient $\operatorname{cov}(K_{-1,n}^{m_1}, X_n^{K_1})/\beta_1^{K_1}\operatorname{var}(X_n^{K_1})$ and $\beta_0^{K_1}$ by passing the line through the means. The factor loading on $K_{-1,n}$ in the second measurement equation, $\alpha_1^{K_{-1,n}}$, is given by $\operatorname{cov}(K_{-1,n}^{m_1}, X_n^{K_1})/\operatorname{var}(X_n^{K_1})$ and the location parameter, $\alpha_0^{K_{-1,n}}$, as before by passing the line through the means. The variance of the unobservable, $\operatorname{var}(\omega_n^{K_1})$, is then derived from the covariance between the two measurements, $\operatorname{cov}(K_{-1,n}^{m_1}, K_{-1,n}^{m_2})$, and the measurement error variances from the variance of the measures. The same identification argument applies to the other endowments.

In addition, consider a nonparametric representation of the latent factors for student and teacher effort, which, consistent with the theory, takes the form

$$\varepsilon_n = f_1(\vec{X}_n, X_i, \omega_n), \qquad (24)$$

$$\varepsilon_t = f_2(\vec{X}_n, X_t, \omega_t), \tag{25}$$

where \vec{X}_n is the set of observable student characteristics of the students in the class, which determine student endowment latent factors; X_i are the observable teacher characteristics that determine teacher endowment latent factors; and to save on notation, we have collapsed school-level, classlevel, and student-level unobserved components of the latent factors of effort into a single unobserved component of the latent factor, ω .²³ Then the measurement system given by (23), (13) and (9), (16) and (11), (24) and (17), and (25) and (18) corresponds to a system of nonlinear measurement equations as given by (3.7) in Cunha et al. (2010). Given the orthogonality assumptions and the parametric forms of the measurement equations along with the distributional assumptions about the measurement errors, we can invoke the identification argument in their theorem 2 to identify the production function parameters.²⁴

In the preceding argument, we did not take advantage of the parametric form of the student and teacher effort functions, (7) and (8). Letting \vec{K}_{-1} and $\vec{\theta}$ be the endowment vector of all students in the class, if we consider the particular measure of student and teacher effort that establishes the units of the latent factor ($\alpha_1^{e_1,m}, \alpha_2^{e_2,m} = 1$), then the effort measurement equations can be written as

$$e_n^m = \underline{e}_s + e_n \Big(\vec{K}_{-1}, \vec{\theta}, a_l, \phi_l; \bar{\delta}_0, \sigma_{\delta_0}^2, \delta_1, \kappa, \gamma_0, \gamma_1, \gamma_2 \Big) + \varsigma_n^{\varepsilon_1}, \qquad (26)$$

$$e_t^m = \underline{e}_{\tau} + e_t \Big(\vec{K}_{-1}, \vec{\theta}, a_t, \phi_t; \bar{\delta}_0, \sigma_{\delta_0}^2, \delta_1, \kappa, \gamma_0, \gamma_1, \gamma_2 \Big) + \varsigma_n^{\varepsilon_2}, \qquad (27)$$

where $e_n(\cdot)$ and $e_i(\cdot)$ are given by (7) and (8), and \underline{e}_s and \underline{e}_τ are the location parameters, $\alpha_0^{\varepsilon_i,m}$ and $\alpha_0^{\varepsilon_i,m}$ in (17) and (18). Given that the only parameters that enter the effort latent functions are those entering the production function, which are identified, and that the latent endowment functions are known, the location parameters can be identified from a single measure of student and teacher effort.

²³ As can be seen from (7) and (8), ω_n and ω_t will generally not be independent as they depend on the same set of student and teacher endowment latent factors.

 $^{^{24}}$ The condition from Cunha et al. (2010) is that there would need to be at least two continuous measures and one additional continuous or discrete measure of student and teacher effort to nonparametrically identify the latent factor functions for student and teacher effort, f_i and f_i .

There are two interpretations of the location parameters. One is simply that they represent reporting bias. If positive (negative), say for student effort, the implication is that students overstate (understate) their effort. A second interpretation is that the location parameters represent effort amounts that are unproductive, for example, time spent doing homework while chatting with or texting friends or while watching television (we provide evidence for this behavior below).

Finally, because the full measurement system includes (22), and (26) and (27), the model imposes cross-equation restrictions; measures of effort provide overidentification of the production function parameters. Indeed, we use the overidentifying restrictions implied by the parametric form of the effort functions to perform a test of the model, which we describe below.

IV. Data

We estimate the model using data on control group students in the tenth grade in the third (and last) year of the ALI program. As part of the ALI experiment, in addition to the curriculum-based end-of-year tests, extensive surveys were administered to both students and teachers that included questions that can serve as measures of the latent factors in the model. We choose the final year of the project because there are more and better measures of the latent variables available in that year. We choose the tenth grade because we have multiple measures of initial mathematics knowledge at that grade level. We estimate the model using data on the measures of latent student and teacher endowments, measures of student and teacher effort, and a measure of end-of-year knowledge. The academic year is divided into two semesters. It is not possible to estimate the model accounting for multiple semesters and for compositional changes within classes. To avoid this problem, we base the estimation on class assignments in the second semester.²⁵

Table 2 provides a list of the variables used in the estimation and a categorization according to their respective latent factors. The determinants of the latent factors (the X's) are background information on students and teachers collected in the surveys. We use data on the latent measures from the second semester (separate surveys were administered at the end of each semester) with respect to both the students and the

²⁵ Most students stay together in both semesters, although some drop out between semesters and the same teacher does not always stay with the same group of students, even when the composition of the class does not change. On average, 76 percent of the students were in the same class in both semesters. In about one-half of the classes, over 90 percent of the students in their second-semester class were also together in the same class in the first semester. In 35 percent of the second-semester classes, all of the students had the same teacher in the first semester. In 38 percent, none of the students had the same teacher as in the first semester.

Latent Variable	Determinants	Measures
Student entering knowledge (K_{0n})	Gender, height, pri- mary school average, parental education, age	9th-grade mathematics ENLACE score, 9th-grade mathematics class grade
Student knowledge preference (θ_n)	Gender, primary school average, pa- rental education	Extent to which likes math generally, hours per week studies nonmath subjects, extent to which does only what is required
Student effort (e_n)		Hours per week studies math, per- centage of time pays attention in class, extent of effort on ALI test, attempts to answer difficult questions on ALI test, frequency skips math
Teacher instructional ability (a_i)	Has education degree, has master's degree, teaching experience	Fraction of students who say that teacher "always" knows subject well, fraction of students who say teacher "always" has control of class
Teacher preference for student knowledge (ϕ_i)	Has education degree, has master's degree, teaching experience	Fraction of students who say that teacher "always" cares that they learn the material, fraction of stu- dents who say teacher "always" cares that they pay attention in class
Teacher effort (e_i)		Hours per week spent planning for classes, hours per week helps stu- dents outside of class, gave only multiple-choice tests, prepared students for ALI test
Student final knowl- edge (K_n)		ALI test score

 TABLE 2

 Measures and Determinants of Latent Variables

teacher.²⁶ As seen in the table, the latent factors of effort and end-of-year knowledge have at least two measures and one determinant.

Table 3 provides descriptive statistics of the latent factor determinants.²⁷ The sample consists of 5,220 students and 56 teachers in 28 schools. As seen in the table, the mean age at entry into tenth grade is 15.9 years;

²⁶ The exceptions are the first three measures of student preferences reported in table 2. These preference measures are based on questions from the survey conducted at the start of the first semester to avoid preferences being contaminated by actual experience in their tenth-grade mathematics class.

²⁷ Missing values for height (217 observations), primary school grade point average (376 observations), and parental education (21 observations) are imputed from regressions. The regression imputation for height is done separately for boys and girls and includes age and dummies for schools. The regression (probit) for parental education includes school dummies. The regression for primary school grade point average includes age, gender, parental education, and school dummies. In each simulation draw of the model, we also draw from the residual variances of these regressions to obtain imputed values.

	Mean	Standard Deviation
Students (5,220):		
Age	15.9	.81
Male (%)	52.2	
Height (cm.)	162.9	10.7
Primary school grade average (6.0–10.0)	8.67	.79
Parent's education less than high school (%)	60.1	
Teachers (56):		
Has education degree (%)	23.2	
Has master's degree (%)	23.2	
Teaching experience (years)	19.5	9.63
Schools (28):		
Class size	33.9	9.21

 TABLE 3

 Descriptive Statistics: Latent Factor Determinants

the modal age is 16, and 95 percent of the students are between the ages of 15 and 17. Slightly over 50 percent of the students are male, their mean grade in primary school (grades 1–6) is 8.7 on a scale from 6 to 10, neither parent attended high school for 60.1 percent of the students, and the students are, on average, 162 centimeters in height. With respect to the teachers, 23 percent have an education degree and the same percentage, though not exactly the same teachers, have a master's degree. The teachers, on average, have almost 20 years of teaching experience. The average class size is 33.9, which is a slight understatement because the sample does not include students who did not complete the survey.

As seen in table 2, the measures of student initial knowledge are the student's (standardized) score on the ninth-grade mathematics ENLACE and the student's grade in his or her ninth-grade mathematics class. The latter is provided in nine categories, at 0.5 intervals beginning at 6.0 and ending at 10.0 (with 10.0 a separate category). We treat it as continuous, coded at midpoints of the intervals, with truncations at 6.0 and 10.0. Teacher instructional ability is measured by student responses to questions about the teacher's subject matter knowledge and control of the class. They are coded according to the fraction of students in the class who rate the teacher as always being in command of the subject and always being in control of the class. The measures of student effort are the number of hours per week the student studied mathematics outside of class $(0, 1, 2, \ldots, 10)$, treated as continuous and truncated at 0 and 10), the percentage of time the student paid attention in class (0-24, 25-49, 50-74, 75-100 percent, treated as ordered categorical), whether or not the student put in a great deal of effort on the ALI test, whether or not the student tried to answer the difficult questions on the ALI test, and the number of days the student skipped the math class. Teacher ef-

fort is measured by the amount of time in hours per week devoted to lesson planning (0–1, 1–2, 2–3, 3–4, 4–6, 7.5, treated as continuous and coded at interval midpoints with truncations at 0 and 7.5), the amount of time in hours per week the teacher tutored the students outside of class (0, 1–3, 4–6, 7–10, 12.5, treated as continuous and coded at interval midpoints with truncations at 0 and 12.5), whether or not the teacher gave only multiple-choice tests, and whether or not the teacher helped students in preparation for the ALI test.

The optimal effort of students and teachers depends on the utility they attach to acquiring knowledge. The student measures of preference for knowledge include the number of hours per week the student studied subjects other than mathematics, the response to a direct question on whether the student likes mathematics (four ordered categories), a measure of conscientiousness, and a self-report of whether the student does only the work that is required (five ordered categories). The teacher's preference, like ability, is measured from the perspective of the students in the class. The two measures are the fraction of students in the class who report that the teacher always is concerned that they learn the material and the fraction who say that the teacher always cares that they pay attention in class.

Descriptive statistics for the measures of the latent factors are presented in table 4. In addition, we present *p*-values for the test that the measures of the same latent variables are pairwise correlated as well as the sign of the correlations.²⁸ Nonindependence of the measures is crucial for identification. In terms of effort, the mean number of hours per week students report that they study math outside of classroom time seems fairly high, 4.68 hours. On the other hand, less than half of the students report that they pay attention in class at least 75 percent of the time. Only a third of the students reported that they put a great deal of effort into the ALI test, but two-thirds said that they attempted the difficult questions on the test. All of the measures have pairwise correlations that are statistically significant, and all pairwise correlations are positive except those that include the number of days class is skipped. With respect to their mathematics knowledge at high school entry, students in these schools scored higher than average on the national ninth-grade ENLACE, 530.9 as opposed to a national average of 500, and received, on average, just over a B in their ninth-grade mathematics course.²⁹ The two measures are positively correlated and the correlation is statistically significant. Among the student preference measures, about two-thirds of the students like math only sometimes,

²⁸ The *p*-values are obtained either from regressions (accounting for clustering at the school level) if one of the measures is continuous (the dependent variable) or from a chi-square statistic based on a contingency table if both variables are discrete.

²⁹ Part of the reason they score higher on the ENLACE is that high school enrollment is not universal while the ENLACE is administered to all ninth-grade students.

	Mean	Standard Deviation
	А	. Students
Measures of student effort (e_n) :		
1. Hours/week studies math	4.68	2.92
2. Percent of time pays attention in class:		
0-24	5.2	
25-49	12.6	
50-74	35.0	
75–100	47.3	
 Puts great deal of effort into ALI test (%) Attempted difficult questions on ALI 	33.1	
test (%)	67.3	
5. Days skipped mathematics class	.41	1.0
p-value: test corr $(1, 2) = 0$, sign		<.001, +
p-value: test corr $(1, 3) = 0$, sign		<.001, +
p-value: test corr $(1, 4) = 0$, sign		<.001, +
p-value: test corr $(1, 5) = 0$, sign		<.001, -
p-value: test corr(2, 3) = 0, sign		<.001, +
<i>p</i> -value: test corr $(2, 4) = 0$, sign		<.001, +
<i>p</i> -value: test corr $(2, 5) = 0$, sign		<.001, +
<i>p</i> -value: test corr $(3, 4) = 0$, sign		<.001, +
<i>p</i> -value: test corr $(3, 5) = 0$, sign		<.001, -
<i>p</i> -value: test corr $(4, 5) = 0$, sign		<.001, -
Measures of student initial knowledge (K_{-1}) :	590.0	104.0
1. 9th-grade ENLACE score	530.9	104.2
2. Mathematics grade in 9th-grade class	8.26	1.03
<i>p</i> -value: test corr $(1, 2) = 0$, sign		<.001. +
Measures of student preference for		
knowledge (θ_n) :		
1. LIKES IIIaui generally (70).	5.9	
Almost never	12.1	
Sometimes	15.1 65.6	
Always	16.1	
9 Does only the work required $(\%)$:	10.1	•••
2. Boes only the work required (70).	6.9	
Somewhat agree	45.1	
Neither agree nor disagree	20.8	
Somewhat disagree	21.4	
Strongly disagree	6.6	
3. Hours/week studies nonmath subjects	5.56	3.04
<i>p</i> -value: test corr $(1, 2) = 0$, sign		<.001, +
<i>p</i> -value: test corr(1, 3) = 0, sign		<.001, +
<i>p</i> -value: test corr(2, 3) = 0, sign		.002, +
	B	. Teachers
Measures of teacher effort (a) .		
1 Hours/week planning lessons	3 99	9 81
9 Hours/week tutors students out of class	3.55 3.11	2.51
3 Addressed student problems in	5.11	4.00
preparation for the ALI test (%)	24.5	
4. Gave only multiple-choice exams (%)	33.9	

 TABLE 4

 Descriptive Statistics: Measures of Latent Variables

	Mean		Standard Deviation
p-value: test corr $(1, 2) = 0$, sign		<.001,+	
<i>p</i> -value: test corr $(1, 3) = 0$, sign		.536, +	
<i>p</i> -value: test corr $(1, 4) = 0$, sign		.541, +	
<i>p</i> -value: test corr $(2, 3) = 0$, sign		.634, +	
<i>p</i> -value: test corr $(2, 4) = 0$, sign		.879, +	
<i>p</i> -value: test corr $(3, 4) = 0$, sign		<.001, +	
Measures of teacher's instructional ability (a_t) :			
1. Fraction of students report teacher			
always in command of subject	.822		.151
2. Fraction of students report teacher			
always in control of class	.614		.203
<i>p</i> -value: test corr $(1, 2) = 0$, sign		<.001, +	
Measures of teacher preference for student			
knowledge (ϕ_t) :			
1. Fraction of students report teacher			
always concerned they learn the material	.783		.165
2. Fraction of students report teacher			
always cares that they pay attention	.700		.176
<i>p</i> -value: test corr $(1, 2) = 0$, sign		<.001,+	

TABLE 4 (Continued)

a little over one-half of the students either strongly or somewhat agree that they do only the work that is required, and, on average, students spend 5.5 hours per week studying nonmath subjects. This latter figure is perhaps surprising in that it is less than 1 hour more than the amount of time spent studying math alone.³⁰ The three measures of preferences are all positively correlated and the correlations are statistically significant.

Teachers report spending almost 4 hours a week preparing lessons and a little over 3 hours a week tutoring students outside of class. The correlation of these measures of teacher effort is positive and statistically significant. The two measures, addressing student problems in preparation for the ALI test and giving only multiple-choice exams, have a positive and statistically significant correlation with each other but do not have statistically significant correlations with the other two measures. Teacher preferences and instructional ability each have two measures, and in each case the two measures are positively correlated and the correlation is statistically significant.

As seen in (7), optimal student effort depends on the student's preference for knowledge and initial mathematics knowledge, the teacher's preference for student knowledge and the teacher's ability, and a composite of the initial knowledge and preference for knowledge of all of

³⁰ These high schools, though now college preparatory, were originally technical high schools and still require that mathematics be taken all 3 years.

the students in the class. Moreover, student effort depends positively on each of these latent factors, including the class composite, although the class composite matters only because students react to teacher effort. Similarly, teacher effort, as seen in (8), depends positively on the same latent factors (except there is no individual student latent preference or initial knowledge). These dependencies imply that student and teacher effort are positively correlated, a direct consequence of the assumption that they are complementary inputs in the production of student knowledge. To be consistent with this prediction, it is necessary that at least some student and teacher effort measures be positively correlated. The correlation between the mean number of hours per week students with a given teacher spend studying math and hours per week the teacher spends preparing lessons is positive with a *p*-value for the test that they are unrelated equal to $.030.^{31}$

We previously argued that the data on student and teacher effort would not support the all-zero effort equilibrium, that is, the equilibrium in which all students in a class exert no effort or the teacher exerts zero effort. To see that, note that only 2.8 percent of the students in the sample report spending 0 hours per week studying math and only 13.1 percent report spending an hour or less. Students seem to be distributed relatively randomly across classes with respect to study time; on average, 2.8 percent of students in a class study 0 hours and 12.9 percent study an hour or less. More relevant is that the largest percentage of students who study 0 (one or less) hours in any of the 154 classes is 14.8 (24.3) percent. Similarly, in terms of teacher effort, only 1.5 percent of teachers report spending less than an hour per week on planning lessons, and none of those also report spending 0 hours tutoring students outside of class.

V. Estimation Results

A. Parameter Estimates

Estimation is carried out with the distributional assumptions previously described. To conserve on parameters, we place some a priori zero covariance restrictions on Λ_2 (school-level unobservables); specifically, the non-zero covariances include the following pairs: (K_{-1}, θ_n) , (K_{-1}, a_t) , and (θ_n, a_t) . In addition, we did not attempt to identify the covariance matrix of class-or teacher-level unobservables to conserve on parameters. Because we set Λ_3 to zero, all between-class variation within schools in teacher instructional ability and preferences and in student initial knowledge and preferences is due to systematic differences in observable student and teacher characteristics.

³¹ Note that there are only 56 tenth-grade math teachers in the 28 schools. The *p*-value is based on a regression of one measure on the other clustered at the school level to account for multiple teachers in a school.

	AND PREDICTED SU	JMMARY STATISTICS	
Parameter	Value	Parameter	Value
$\overline{\delta_0}$.937	γ_1	.372
σ_{δ}	.055	γ_{11}	.00017
κ	.0066	γ_2	.082
δ_1	.011	$\gamma_{\scriptscriptstyle 21}$	0016
γ_0	.549		
	Predicted Sum	nmary Statistics ^a	
Endow	ments	Outc	omes
μ_{K-1}	528.5	μ_K	498.0
σ_{K-1}	47.7	σ_{K}	52.6
$\mu_{ heta n}$	5.61	μ_{en}	3.14
$\sigma_{\theta n}$	2.06	σ_{en}	.939
μ_{at}	.830	$\mu_{\scriptscriptstyle et}$	1.59
σ_{at}	.136	σ_{et}	.523
$\mu_{\theta t}$.794		
$\sigma_{\theta t}$.156		

 TABLE 5

 Production Function Parameter Estimates

 and Predicted Summary Statistics

^a Based on 5,000 simulations of each class.

The model contains 85 parameters. Table 5 presents estimates of the production function parameters along with selected summary statistics of the predicted latent factors.³² The parameters indicate that end-of-year knowledge increases with initial knowledge ($\delta_0 > 0$). However, because the measures of K_n and $K_{-1,n}$ are not cardinal measures of a single scale of mathematics knowledge, δ_0 reflects both knowledge depreciation and a normalization. The marginal product of student effort, γ_1 , is positive and is essentially independent of class size ($\gamma_{11} \approx 0$). The marginal product of teacher effort, γ_2 , is also positive and declines with class size ($\gamma_{21} < 0$); at the largest class size the marginal product of teacher effort is close to zero. The normalizations (δ_0 , κ) are statistically significant at conventional levels as is the student's marginal product of effort (γ_1). However, neither the teacher's marginal product, the coefficient on teacher's ability (γ_0), nor the effects of class size are statistically significant.

Mean student effort is equivalent to 3.14 hours per week of math study time and mean teacher effort to 1.59 hours per week of class preparation time. The coefficient of variation is similar for student and teacher effort, 0.30 versus 0.33. The amount of effort that is unproductive (or subject to reporting bias) is large for both students and teachers. The unproductive (or overstated) effort of students (\underline{e}_s) is the equivalent of 1.61 hours per week of study time, about 35 percent of "total" effort. Similarly, the

³² The entire set of parameters and the standard errors are shown in app. table A1.

unproductive (or overstated) effort of teachers (\underline{e}_t) is the equivalent of 2.51 hours per week of class preparation time, about 60 percent of "total" effort.

Mean end-of-year knowledge (K_n) is equivalent to a standardized score of 498.0 on the ALI test, with a standard deviation of 52.6.³³ Mean initial knowledge is equivalent to a (standardized) score of 528.5 on the ninthgrade ENLACE, with a standard deviation of 47.7.³⁴ The coefficient of variation of student preferences for knowledge is greater than that of teacher preferences for student knowledge, 0.40 versus 0.20. Teacher ability has the smallest coefficient of variation, 0.16.

B. Model Fit

Table 6 compares actual and predicted statistics for each latent variable measure. It also shows the fraction of the measure's variance that reflects variation in the latent factor (one minus the fraction due to measurement error). The measures are categorized according to their corresponding latent. The first row shows the fit for the ALI test, the measure of end-of-year (tenth-grade) math knowledge (K_n) . As seen, the model fits the mean and standard deviation well. According to the model estimates, 27.8 percent of the variance in the ALI test is accounted for by the variation in math knowledge (and the residual by measurement error). In comparison, 20.3 and 31.5 percent of the variance of the two measures of initial knowledge (K_{0n}) , the ninth-grade ENLACE score and the students' grades in their ninth-grade mathematics class, is accounted for by the variation in initial math knowledge.³⁵

The fit for the measures of the other latent factors is generally quite good, although measures vary in their degree of measurement error. For example, the actual mean hours per week of study time devoted to mathematics, one of the measures of student effort, is 4.68, while the predicted mean is 4.79 hours; the actual standard deviation of the measure is 2.92 hours and the prediction is 2.96 hours. Hours of study time is, however, a noisy measure of student effort, with the noise component of the measure accounting for 92 percent of its variance. Another measure of student effort, the percentage of time the student reports paying attention in class, which

³³ On the basis of an analysis of answer sheets, Behrman et al. (2015) report that 3.7 percent of students in the control group engaged in copying. We account for this in our estimation by including an indicator variable in the test score measurement equation for whether a student was identified as a copier. Our estimate indicates that copiers increased their test scores by 22.1 standardized points on average.

³⁴ In terms of the second measure of initial knowledge, mean initial knowledge is equivalent to a grade in the ninth-year mathematics course of 8.23.

³⁵ Regressions of the ENLACE score and the ninth-grade mathematics class grade on their observable determinants (listed in table 1) yield R^2 's of 11 and 24 percent. The rest of the variance in initial knowledge is due to the unobserved component.

	Аст	UAL	Pred	ICTED	
	μ	σ	μ	σ	$\sigma_{\text{True}}^2/\sigma_{\text{Total}}^2$
Measure of 10th-year math knowledge (K) :					
ALI test	500.5	99.8	498.0	100.1	.278
Measures of student effort (e_n) :					
Hours/week studies math	4.68	2.92	4.79	2.95	.078
Days attended math class	79.6	1.10	79.6	1.06	.150
Percent of time pays attention in class:					.148
0-24	5.3		5.4		
25-49	12.6		12.7		
50-74	35.0		34.9		
75–100	47.3		47.1		
Puts lot of effort into ALI test (%)	33.1		33.0		.062
Tried to answer difficult questions on					
ALI test (%)	67.3		67.1		.065
Measures of teacher effort:					
Hours/week of class preparation	3.99	2.31	4.08	2.34	.031
Hours/week tutor outside of class	3.05	2.94	3.25	3.66	<.001
Gave only multiple-choice tests (%)	37.4		35.8		.034
Prepared students for ALI test (%)	21.3		23.4		.001
Measures of student initial knowledge (K_0) :					
9th-grade ENLACE standardized score	530.9	104.2	528.5	105.9	.203
9th-grade mathematics grade	8.26	1.03	8.22	1.04	.315
Measures of student preference for			~ ~ ~ ~ ~		
knowledge (θ_n) :					
Hours/week studies nonmath subjects	5.56	3.04	5.47	3.32	.242
Likes math (%):					.134
Never	5.2		5.5		
Almost never	13.1		13.2		
Sometimes	65.6		65.3		
Always	16.1		16.0		
Student does only what is expected:					.022
Strongly agree	6.2		6.2		
Agree	45.1		46.2		
Neither agree or disagree	20.8		20.5		
Disagree	21.4		20.9		
Strongly disagree	6.6		63		
Measures of teacher preferences for	0.0		0.0		
student knowledge (θ_i) :					
Fraction of students in class who					
respond that:					
Teacher always cares if they learn	779	179	784	158	806
Teacher always wants students to pay	.115	.174	.704	.150	.000
attention in class	603	170	707	189	825
Measures of teacher instructional	.055	.175	.707	.102	055
ability(a)					
Eraction of students in class who reasond					
that					
Teacher always knows the subject	818	150	818	149	781
Teacher always has control of class	.608	.204	.629	.207	.686

TABLE 6 Model Fit

can take on values within the ranges 0–24 percent, 25–49 percent, 50– 74 percent, and 75–100 percent, also fits well; the actual percentages are 5.3, 12.6, 35.0, and 47.3 while the predicted percentages are 5.4, 12.7, 34.9, and 47.1. The measure is less noisy than hours of study time, with the variance component of the latent factor accounting for 14.5 percent of the total variance of the measure.³⁶ The fit for the measures of teacher effort is similarly good, though the noise components of the measures are larger. The least noisy measure of teacher effort is the number of hours per week spent on class preparation, for which the noise component is 96.9 percent of the total variance. With respect to student preferences, the variance components of the factors for the three measures account for 24.2, 13.4, and 2.2 percent, while for the two measures of teacher preferences, they are 80.6 and 83.5 percent. Variance components of the factors for the measures of teacher instructional ability, based on student reports as was also the case for teacher preferences, account for 73.1 and 68.6 percent.

Further evidence on model fit is provided in table 7, which compares regressions of predicted model outcomes on exogenous student and teacher characteristics that determine the latent factors to regressions of measures of those outcomes on the same variables. Given the nonlinear nature of the production function and the effort functions, which are derived as the solution to the game, the ability of the model to capture these relationships is a demanding test.³⁷ The first two columns show the regression coefficients for the actual ALI test score and the test score predicted from the model. The column based on the actual data also reports the 95 confidence interval and the *p*-value. With respect to student characteristics, three of the five coefficients (age, height, and primary school average) based on the predicted test score fall within the 95 percent confidence interval, which in each case is quite tight (*p*-value $\leq .001$). Two coefficients (male and parent education) also fall within the 95 percent confidence interval, but the *p*-values exceed .05. With respect to teacher characteristics, the regression coefficients based on the predicted test score fall within the 95 percent confidence interval of those based on the actual data for all three characteristics, but p-values are large.

The model is less successful with respect to the student effort measure, hours per week devoted to math study time. In the two cases in which the p-value of the regression coefficient based on the actual data was less than .05, the regression coefficient based on the prediction fell inside the

³⁶ Measures that are ordered categorical or binary are treated as themselves coming from an underlying continuous latent variable. The proportion of variance that the measure explains of the model latent factor (in this case the preference for knowledge, θ_n) is taken with respect to the continuous latent factor of the measure.

³⁷ Note that although the model outcomes are functions of the student and teacher observables, at both the individual and class levels, the functions are highly nonlinear. These misspecified regressions are used only to assess model fit and have no interpretation.

95 percent confidence interval for one of them. For all of the other variables, *p*-values were too large for the comparison to be meaningful, as was also the case for the measure of teacher effort, hours per week spent planning.

C. Discussion

1. Characteristics of the Knowledge Production Function

Table 8 summarizes quantitatively the features of the production function for end-of-year knowledge. Each row shows the effect on the normalized and raw score equivalent level of end-of-year knowledge of ceteris paribus changes in each of the latent factor determinants of end-of-year knowledge. When each latent factor is varied, the others are held constant at mean values. As seen in the table, an increase in initial knowledge from 2 SD below to 2 SD above the mean would increase tenth-grade math knowledge by 180.4 standardized points or by 20.9 raw score percentage points, an increase of 3.4 SD. Although these changes are substantial, even students with initial knowledge 2 SD above the mean would answer, on average, only 48.6 percent of the questions correctly.

Ceteris paribus changes in student and teacher effort or teacher ability have negligible effects on end-of-year knowledge. A student who increases effort from 0 to 2 SD above the mean, equivalent to an increase of 5.28 hours spent studying per week, would increase knowledge by only 0.10 SD. Moreover, the concavity of the production function in student effort is such that three-fifths of the gain in knowledge arises from increasing effort from zero to only 1 SD below the mean. The marginal product of student effort increases with student initial knowledge, but not greatly. A student with initial knowledge 2 SD above the mean who increases effort from 0 to 2 SD above the mean would increase knowledge by 0.12 SD, only 0.02 SD more than a student with initial knowledge at the mean. The results are similar both for teacher effort and for teacher ability. In all cases, the raw score equivalent never reaches 50 percent.

2. Accounting for Performance

The results in table 8 do not provide a complete picture as to the causes of low performance. Student and teacher effort are choices, and changes in student initial knowledge or in teacher ability will affect those choices. To account fully for low performance, we perform a series of counterfactuals in which we change the composition of the classes in terms of student and teacher endowments. In each case, we set a student or teacher endowment for all students or teachers in the sample to be 2 SD above the mean. The results are reported in table 9.

	Mon	TABLE	7			
	MUDE	L II: EXPLANA	I UKY VAKIABLES			
	End-of-Year Kn ALI Test S	IOWLEDGE: SCORE	Student Effort Hours Studi	r Measure: es Math	Teacher Effor Hours Pla	et Measure: anning
	$Data^{a}$	Model	$\operatorname{Data}^{\mathrm{b}}$	Model	$Data^{c}$	Model
Student characteristics:						
Age	-8.71	-6.71*	044	025		
)	(-12.1, -5.31)		(154, .065)			
	<.001		.427			
Female	-5.74	-9.26	255	035*	•	
	(-12.1, .600)		(506,003)			
	.074		.047			
Height	.866	.438*	.0024	.0017		
)	(.421, 1.31)		(010, .015)			
	< 001		.709			
Primary school average	42.4	37.4^{*}	.219	.653 * *		
	(36.7, 48.2)		(.082, .357)			
	<.001		.002			
Parent education less than high school	4.59	2.26	960.	.340		
	(-2.78, 12.0)		(120, .312)			
	.212		.384			

This content downloaded from 128.042.192.005 on December 10, 2018 12:49:19 PM All use subject to University of Chicago Press Terms and Conditions (http://www.journals.uchicago.edu/t-and-c).

las education degree	-6.10	.446	181	.027	251	.206
	-28.6, 16.0)		(500, .137)		(-1.94, 1.43)	
	.576		.265		.766	
Has master's degree	365	.065	.041	.049	.309	000.
	-25.3, 24.6)		(280, .362)		(-1.75, 2.37)	
	.976		.804		.765	
eaching experience (years)	.587	001	.0008	004	.021	008
	562, 1.74)		(014, .016)		(060, .103)	
	.304		.918		.606	
ervations	4,638		4,607		56	

• Tobit coefficients (lower bound is 0.5, upper bound is 7.5), 95 percent confidence interval, and pvalue. Accounts for clustering at the class level.

* The parameter falls within the 95 percent confidence interval and the p-value of parameter < 05. ** The parameter falls outside the 95 percent confidence interval and the p-value of parameter < 05.

	Knowledge: Normalized Equivalent (SD = 53.0)	Knowledge: Raw Score Equivalent (SD = 6.12)
Initial knowledge (K_{-1}) :		
$\mu - 2\sigma = 433$	409.4	27.7
$\mu = 529$	499.6	38.2
$\mu + 2\sigma = 624$	589.8	48.6
Student effort (e_n) :		
Zero	495.0	37.6
$\mu - 2\sigma = 1.27$	498.3	38.0
$\mu = 3.14$	499.6	38.2
$\mu + 2\sigma = 5.02$	500.5	38.3
Student effort (e_n) at $K_{-1} = 624$:		
Zero	584.4	47.9
$\mu - 2\sigma = 1.27$	588.2	48.4
$\mu = 3.14$	589.8	48.6
$\mu + 2\sigma = 5.02$	590.8	48.7
Teacher effort (e_i) :		
Zero	495.0	37.6
$\mu - 2\sigma = 1.27$	499.5	38.1
$\mu = 3.14$	499.6	38.2
$\mu + 2\sigma = 5.02$	499.7	38.2
Teacher ability (a_i) :		
Zero	495.0	37.6
$\mu - 2\sigma = .56$	499.2	38.1
$\mu = .89$	500.0	38.2
$\mu + 2\sigma = 1.10$	500.4	38.2

 TABLE 8

 Properties of the Knowledge Production Function: Predicted Measures

 of Performance at Alternative Values of Student and Teacher Inputs

The first column shows the baseline mean student and teacher effort levels and student end-of-year knowledge (both standardized and raw score). The second column shows those same statistics for the case in which initial knowledge is 2 SD above the mean for all of the students in each class. As seen, student effort increases by the equivalent of 0.31 hour per week of study time and teacher effort by 0.17 hour per week of class preparation time. The reason for these increases, though small, stems from the increase in the marginal product of effort with increasing initial knowledge. End-of-year knowledge increases by 90.1 standardized points (1.7 SD). However, the raw score equivalent remains under 50 percent. The second column reports the same outcomes for the case in which instructional ability is 2 SD above the mean for all teachers. The effect on student and teacher effort is similar to that for the change in student initial knowledge. However, because the marginal product of teacher ability is also low, student knowledge increases only by 1.1 standardized points. Thus, improving teacher instructional ability, at least within the range of abilities in the data, is not a viable mechanism to improve student knowledge of the current curriculum.

	Prefer	ences on Tenth	-GRADE MATHEM	IATICS KNOWLED	GE
	Baseline	$K_{-1} = \mu + 2\sigma$	$a_t = \mu + 2\sigma$	$\theta_n = \mu + 2\sigma$	$\theta_t = \mu + 2\sigma$
e_n	3.14	3.45 (.33 SD)	3.48 (.33 SD)	4.46 (1.41 SD)) 3.15 (.01 SD)
e_t	1.59	1.76 (.33 SD)	1.75 (.33 SD)	1.71 (.23 SD)	1.88 (.55 SD)
K_n :					
Standard					
score	498.0	588.1 (1.7 SD)	499.1 (.01 SD)	498.7 (.01 SD)	498.0 (<.001 SD)
Raw score					
(%)	38.0	48.4 (1.7 SD)	38.1 (.01 SD)	38.1 (.01 SD)	38.0 (<.001 SD)

TABLE 9
THE EFFECT OF UNIVERSAL CHANGES IN STUDENT INITIAL KNOWLEDGE,
TEACHER INSTRUCTIONAL ABILITY, AND STUDENT AND TEACHER
Preferences on Tenth-Grade Mathematics Knowledge

The counterfactual simulations related to changing preferences (more accurately to the ratio of the marginal utility of knowledge to the per-unit cost of effort) also lead to small effects on end-of-year knowledge. Unlike student initial knowledge and teacher ability, the effect of changing preferences on knowledge arises only through the effect on effort. As seen, setting student preferences at 2 SD above the mean increases average student effort by the equivalent of 1.32 hours of study time per week and average teacher preparation time by the equivalent of 0.12 hour. Even with these increases in effort, however, average end-of-year knowledge increases by only 0.7 standardized points (0.01 SD). The reason that students exert as much effort as they do, despite the low productivity of effort, is that they value even small increments of knowledge.³⁸ Setting teacher preferences at 2 SD above the mean has essentially no effect on average student effort and increases average teacher effort by 0.29 hour per week of class preparation time. The overall effect on student knowledge is again essentially nil.

The counterfactuals in tables 7 and 8 demonstrate that (1) the initial preparation of students in terms of their incoming math knowledge has a large impact on tenth-grade math performance as measured by the standardized score but still leaves even those students with the best preparation having mastered less than 50 percent of the curriculum; (2) within the range of instructional abilities and preferences of teachers observed in the data, employing "better" teachers would have a small impact on student end-of-year curriculum knowledge; and (3) increasing the enjoyment that students receive from acquiring math knowledge has a considerable impact on student effort and a smaller impact on teacher effort but induces only a small increase in student end-of-year knowledge.

As Cunha et al. discuss (2010), a limitation of test score analyses is that they generally identify only determinants of relative performance. The

³⁸ One possibility is that small increases in this measure of knowledge have a large effect on the student's course grade, and it is the latter that students care most about and that matter in terms of future academic success.

ALI test was designed to uncover the extent to which students have mastered the curriculum, and as such, raw scores (percentages correct), which our analysis focuses on, are a measure of their mastery. Nevertheless, as suggested by Cunha et al., it would be desirable to know whether performance on the ALI test is related to other outcomes of interest, such as attending college or adult earnings. Unfortunately, the ALI experiment did not follow students subsequent to leaving high school.³⁹

3. Credibility and Interpretation of Results

In this section, we raise two questions: (1) Are the results credible? and (2) What is the interpretation of the results?

Credibility.—Obviously, the estimation results rely on many assumptions: strategic interaction between students and teachers, specific functional forms of the production function and preferences, specific distributional assumptions, covariance and exclusion restrictions, and the match of measures to latent factors. It is not possible to test each of these assumptions separately. Yet the credibility of the results depends on the ability of the model to capture "true" systematic aspects of student and teacher effort decisions.

As we have shown, the model places cross-equation restrictions between the parameters of the knowledge production function and the derived effort functions. Such restrictions could have led to a poor fit of the estimated model to the data. That did not, however, seem to be the case.⁴⁰ A more convincing validation exercise would be to directly test the crossequation restrictions. We carried out such a test by relaxing the restriction on three of the parameters, δ_0 , γ_1 , and γ_2 , estimating separate parameters for the production function and for the effort functions. We did not estimate separate parameters for κ , which is a normalization that is identified only in the production function. Also, because teacher ability and teacher preferences are highly correlated, we did not estimate a separate parameter for γ_0 , which would be difficult to distinguish from γ_1 and γ_2 in the effort functions. In essence, the model without the restrictions breaks the tight link between fitting the measure of end-of-year knowledge (the

⁴⁰ However, as argued in Wolpin (2007), model validation based on within-sample fit can be problematic because of structural data mining,

³⁹ Cawley, Heckman, and Vytlacil (2001) summarize extensive empirical evidence showing that cognitive skills strongly predict schooling attainment and wages. Heckman, Stixrud, and Urzua (2016) show that cognitive and noncognitive skills also explain a variety of other labor market and behavioral outcomes, including employment, work experience, choice of occupation, teenage pregnancy, smoking, and criminal participation. Much of the accumulated evidence on the relationship between cognitive and noncognitive skills and adult outcomes is based on National Longitudinal Survey of Youth 1979 survey data, which gathered extensive test score data on teenage youths and followed them for more then 20 years, allowing examination of subsequent outcomes.

ALI test score) and the measures of student and teacher effort. The variation in student effort, for example, that must be attributed to preference variation to explain the level of student effort, while accommodating the low marginal product of effort, can now be explained by changing the effort function parameters. We reestimated the model allowing for the separate parameters and performed a likelihood ratio test. We did not reject the hypothesis that the restrictions were valid.⁴¹

As we noted, model restrictions allowed us to identify the amount of effort that is unproductive, estimated for students to be 1.6 hours per week of study time and for teachers to be 2.5 hours per week of preparation time. In addition to study time, students were also asked to report on competing activities they engaged in while doing homework. Almost 30 percent of the students reported that they always engage in one of the following activities while doing homework: chatting or texting, watching television, playing video games, or caring for a sibling.⁴² If we assume that productive study time is zero for those students, the average number of productive hours of study time would be 3.5. Recall that our estimate of productive effort is 3.1 hours per week.

Interpretation.—The results are rather dramatic: the only factor that really matters in determining how much of the tenth-grade mathematics curriculum was mastered is mathematics knowledge upon entry into the tenth grade. Moreover, even students who enter with knowledge acquired through the ninth grade that is 2 SD above the mean (the equivalent of a grade of 9.3 in their ninth-grade mathematics class) master less than 50 percent of the tenth-grade curriculum. We explore several possible explanations.

One explanation is that students may put little or no effort into the ALI test, because the test is perceived as low stakes. Although we cannot rule out the possibility that test-taking effort is partly responsible for the results, it is unlikely to be the main explanation. First, the difference between the average raw score, 38 percent, and the raw score that would be achieved by guessing, 25 percent, is statistically significant. Second, the vast majority, 89 percent, of enrolled students took the test. This figure is comparable to that of the treatment group, 92.5 percent, students who had a monetary incentive to sit for the test. Third, as we have already noted, 67 percent of the control group students reported that they attempted to answer the difficult questions on the test. In the treatment group, the figure was 81 percent. Finally, the relevant question is a counterfactual; namely, among students who put in less than "full" test-taking

⁴¹ Twice the difference in the log likelihood values of the unrestricted and restricted models was 7.33. The chi-square statistic with 3 degrees of freedom is 7.81 at the .05 critical value.

 $^{^{\}rm 42}$ The response categories for each of them were never, almost never, sometimes, and always.

effort, what would their performance have been had they put in "full" effort? Although we cannot give a quantitative answer to that question, we can provide some information about reported test-taking effort and student characteristics. In particular, the students reporting that they attempted the difficult questions scored 0.25 SD higher on the ninth-grade ENLACE, reported spending 0.74 hour more per week studying math, and were half as likely to report paying attention in class less than 50 percent of the time than those who reported that they did not attempt the difficult questions. This evidence suggests that those exerting less test-taking effort would have mastered less of the curriculum regardless of their test-taking effort given that they had lower initial knowledge, studied less, and paid attention less during the year.

A second explanation is that the tenth-grade ALI test was just very difficult, and so it is no surprise that test scores were low. It is true that the test was difficult, but it was difficult because the curriculum is difficult. The creators of the test, as noted by a national testing organization, took as their mandate the design of a test that was faithful to the curriculum. Thus, our estimate of tenth-grade math knowledge, based on a raw score equivalent metric, can be taken as a measure of the fraction of the curriculum that was mastered. Mastering the curriculum is presumably the goal of the Ministry of Education.

A third explanation is that teachers do not follow the curriculum. Obviously, students cannot learn what they are not given to learn. Teacher reports, however, indicate that the curriculum is followed. Eighty percent of the teachers report that they taught at least 75 percent of the trigonometry curriculum, sequentially the last and most difficult component of the tenth-grade curriculum.⁴³ And, of course, the motivation for not teaching the curriculum may be that the curriculum is not learnable (teachable) given the level of mathematics knowledge with which students enter.

The final explanation, and to us the most convincing, is that the tenthgrade curriculum is too difficult given the level of mathematics knowledge acquired through the ninth grade. To be consistent with our results, it would have to be true that even those students far above the average in the distribution of the latent factor of initial knowledge do not have the requisite foundation for learning the material covered in the tenth-grade curriculum, and further, at least within the range of the data, no amount of effort on the part of the student or teacher and no level of teacher ability would compensate for the deficiency. Given the institutional structure, it is reasonable to assume that the ninth- and tenth-grade mathematics curricula are aligned, that is, that mastery of the ninth-grade curriculum would

⁴³ The raw score for the trigonometry component of the test was only 29 percent. The raw scores for algebra and geometry were both about 40 percent.

provide an adequate foundation.⁴⁴ Under that assumption, the interpretation of our results depends on what our measures of initial knowledge, the ninth-grade ENLACE, and the grade earned in ninth-grade mathematics actually indicate about entering mathematics knowledge. An assumption, consistent with our results, is that the initial knowledge latent factor that we identify on the basis of those measures, although capturing some relevant variation among students, does not span the knowledge necessary to successfully tackle the tenth-grade curriculum.⁴⁵

VI. Conclusions

This paper developed and estimated a strategic model of the joint effort decisions of students and teachers in a classroom setting to understand the reasons for low mathematics performance of Mexican high school students on curriculum-based examinations. The model allowed for student and teacher heterogeneity in preferences for knowledge acquisition and in student initial mathematics preparation and teacher instructional ability.

Survey data of students and teachers collected as part of the ALI project provided multiple measures of student and teacher effort, student and teacher preferences, student initial knowledge, and teacher ability, all of which were treated as latent variables. An end-of-year curriculum-based test provided a measure of tenth-year mathematics knowledge. A simulationbased maximum likelihood estimation procedure was used to recover the parameters of the knowledge production function as well as parameters governing the latent variable and measurement structures.

Estimation results indicated that the most important factor accounting for low performance is the lack of sufficient prior mathematics preparation. On the basis of the production function estimates, a ceteris paribus increase in student effort from its mean to 2 SD above the mean, an increase equivalent to almost 2 extra hours per week of time devoted to studying math, was estimated to increase end-of-year knowledge by only 0.02 SD. A similar increase in teacher effort increased knowledge by less than 0.01 SD. In contrast, a 1 SD increase in student initial knowledge increased end-of-year knowledge by 0.85 SD.

⁴⁴ In both the ninth and tenth grades, the curriculum is set at the federal level, although the states are responsible for oversight of lower secondary schools (grades 7–9). Our sample of tenth graders comes from schools for which the federal Ministry of Education has oversight.

⁴⁵ Unfortunately, we do not have a mapping from the standardized score on the ninthgrade ENLACE to a raw score as we do for the tenth-grade ALI test. Thus, we cannot determine how much of the ninth-grade curriculum is mastered. However, looking at the ENLACE test, it appears that it has a level of difficulty similar to that in the tenth-grade ALI test. And only 11 percent of the students in our sample were deemed proficient on the basis of the ENLACE.

We also used the model to perform counterfactual experiments that changed the composition of classes in terms of student and teacher endowments, incorporating the implied optimal changes in student and teacher effort. Increasing student initial knowledge in all classes to be 2 SD above the sample mean increased end-of-year knowledge by 1.7 SD above the mean. However, this seemingly large effect corresponded to mastery of only 48.4 percent of the curriculum. On the other hand, similarsize changes in student or teacher preferences or in teacher ability led to negligible effects on end-of-year knowledge.

We considered a number of explanations for these results. The most compelling in our view is that there is a mismatch between the tenthgrade mathematics curriculum and the mathematics preparation of incoming high school students. Because of this mismatch, increasing effort per se of either students or teachers leads to only small increases in end-of-year knowledge. Simply having a rigorous curriculum without a strong foundation does not by itself improve knowledge.

	P	arameter Es	TABLE A1 stimates and Standard Errors		
Parameter	Value	Standard Error	Parameter	Value	Standard Error
			Production Function	u	
δ1 Φ5	.937.055	.072 .073	بر ۲۳	.372 .00017	.186 .023
δ11 κ Υ ₀	.011 .0066 .549	.024 .0018 1.33	γ_2^2 γ_{21}	.0819 0016	1.29
			Endowments: Standard Deviation (σ)	and Correlations (ρ)	
School level:			School level:		
$\sigma \xi^{h_0}$	18.7	6.14	$\rho(\xi^{k0},\xi^{\theta n})$.430	609.
$\sigma \xi^{at}$.126	.119	$\rho(\xi^{N}, \xi^{\alpha})$	018	.577
$\sigma \xi^{\theta_n}$.837 149	.295 100	$\rho\left(\xi^{at}, \xi^{\theta t}\right)$.996	.096
student level:	71.1.	COL	Student level:		
ይ ^{&0} ይ ድ	$23.2 \\ 1.40$	5.21.019	$\operatorname{cov}(\xi^{h0}, \xi^{\theta n})$	666.	.002
			Measurements		
<i>K</i> ₀ : initial knowledge: ENLACE score:			<i>a</i> ; teacher ability: Teacher always knows the subject:		
Constant	00.		Constant	0.	
Slope	1.00	:	Slope	1.0	:
Variance	8,950	2,132	Variance	.0069	.012
Math 9th-year grade:			Teacher always has control of class		
Constant	1.27	.730	Constant	452	1.20

Appendix

		T/	ABLE A1 (Continued)		
		Standard			Standard
Parameter	Value	Error	Parameter	Value	Error
Slope	.0132	.001	Slope	1.30	1.39
Variance	.860	.272	Variance	.014	.044
θ_n : student preferences			θ_i ; teacher preferences:		
Likes math:			Teacher always cares if student learns:		
Constant	00.		Constant	0.	
Slope	.191	.189	Slope	1.00	:
Cutoff 1	647	1.12	Variance	.0059	.019
Cutoff 2 – cutoff 1	.764	.519			
Cutoff 3 – cutoff 2	2.02	.530			
Variance	1.00				
Hours/week studies nonmath subjects:			Teacher wants students to pay attention:		
Constant	00.		Constant	176	.509
Slope	1.00		Slope	1.12	.544
Variance	13.3	7.86	Variance	.0061	.020
Student does only what is expected:					
Constant	0.00				
Slope	.073	.159			
Cutoff 1	-1.15'	1.09			
Cutoff 2 – cutoff 1	1.62	.597			
Cutoff 3 – cutoff 2	.555	.268			
Cutoff 4 – cutoff 3	.936	.497			
Variance	1.00				

BIF AL Co

\$\vec{e}_n\$: student effort: Hours/week studies math:			<i>ci</i> ; teacher effort Hours/week of class preparation:		
÷ گ	1.61	1.35		2.51	18.1
Slope	1.00	•	Slope	1.00	•
Variance	10.4	5.37	Variance	8.46	52.7
Days attends math class:			Hours/week tutoring outside of class:		
Ćonstant	79.6	7.39	Constant	2.00	0.27
Slope	.712	1.49	Slope	.0005	.004
Variance	12.2	8.95	Variance	29.4	274
Percent of time pays attention in class:			Gave only multiple-choice tests (Y or N):		
Constant	0.00		Constant	1.10	55.4
Slope	.445	.335	Slope	358	13.2
Cutoff 1	.366	1.45	Variance	1.00	
Cutoff 2 – cutoff 1	.760	.454			
Cutoff 3 – cutoff 2	1.07	.440			
Put much effort into ALI exam (Y or N):			Answered student questions: ALI test (Y or N):		
Constant	-1.76	2.08	Constant	532	49.3
Slope	.274	.420	Slope	047	11.3
Variance	1.00		Variance	1.00	
Attempted difficultALI questions (YorN):					
Constant	878	2.37			
Slope	.281	.521			
Variance	1.00				
T: ALI test score:					
Constant	00.				
Slope	1.00	:			
Variance	7,181	2,766			

		TA	BLE Al (Continued)		
		Standard			Standard
Parameter	Value	Error	Parameter	Value	Error
			Determinants $(X$'s)		
K_0 : initial knowledge:			a_i : teacher ability:		
Gender	-9.65	3.88	Has education degree	.0011	.238
Height	.454	.229	Has master's degree	.023	.270
Age	-6.80	2.74	Years teaching	.014	.014
Primary school GPA	39.3	5.81	Years teaching squared	00047	.00048
Parents' education less than high school	2.22	1.50	Constant	.790	.256
Constant	226.2	78.2			
θ_n : student preferences:			θ_i ; teacher preferences:		
Gender	0009	.0026	Has education degree	.011	.251
Primary school GPA	1.39	.068	Has master's degree	.044	.240
Parents; education less than high school	.910	.141	Years teaching	.013	600.
Constant	-7.02	.705	Years teaching squared	00048	.00033
			Constant	.766	.240
T: ALI test score:					
Copier	22.1	108			

References

- Ackerberg, Daniel A., and Gautam Gowrisankaran. 2006. "Quantifying Equilibrium Network Externalities in the ACH Banking Industry." *RAND J. Econ.* 27 (3): 738–61.
- Austen-Smith, David, and Roland G. Fryer. 2005. "An Economic Analysis of Acting White." Q. I.E. 120 (2): 551–83.
- Bagwell, Kyle. 1995. "Commitment and Observability in Games." Games and Econ. Behavior 8 (2): 271–80.
- Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2015. "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools." *J.P.E.* 123:325–64.
- Ben-Porath, Yoram. 1967. "The Production of Human Capital and the Life Cycle of Earnings." *J.P.E.* 75:352–65.
- Brock, William A., and Steven N. Durlauf. 2001. "Interactions-Based Models." In Handbook of Econometrics, vol. 5, edited by James J. Heckman and Edward Leamer, chap. 54. Amsterdam: North-Holland.
- Cawley, John, James J. Heckman, and Edward Vytlacil. 2001. "Three Observations on Wages and Measured Cognitive Ability." *Labour Econ.* 8 (4): 419–42.
- Cunha, Flavio, and James J. Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." J. Human Resources 43:738–82.
- Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78:883–931.
- De Paula, Aureo. 2009. "Inferences in a Synchronization Game with Social Interactions." J. Econometrics 148:56–71.
- Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2008. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." Working Paper no. 14475, NBER, Cambridge, MA.
- Duflo, Esther, Rema Hanna, and Stephen Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *A.E.R.* 102:1241–78.
- Fruehwirth, Jane Cooley. 2014. "Can Achievement Peer Effect Estimates Inform Policy? A View from Inside the Black Box." *Rev. Econ. and Statis.* 96 (3): 514–23.
- Heckman, James J., Jora Stixrud, and Sergio Urzua. 2016. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *J. Labor Econ.* 24 (3): 411–82.
- Joreskog, Karl G., and Arthur S. Goldberger. 1975. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." J. American Statis. Assoc. 70:631–39.
- Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." Working Paper no. 10971, NBER, Cambridge, MA.
- Lazear, Edward. 2001. "Educational Production." Q. J.E. 116 (3): 777-803.
- Pritchett, Lant, and Amanda Beatty. 2012. "The Negative Consequences of Overambitious Curricula in Developing Counties." Working Paper no. 293, Center Global Development, Washington, DC.
- Sweeting, Andrew. 2009. "The Strategic Timing Incentives of Commercial Radio Stations: An Empirical Analysis Using Multiple Equilibria." *RAND J. Econ.* 40 (4): 710–42.
- Todd, Petra, and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Econ. J.* 113:F3–F33.

— 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *A.E.R.* 94:1384–1417.

Vives, Xavier. 2005. "Complementarities and Games: New Developments." J. Econ. Literature 43:437–79.

Wolpin, Kenneth I. 2007. "Ex Ante Policy Evaluation, Structural Estimation, and Model Selection." A.E.R. Papers and Proc. 97:48–52.