Why Don't Struggling Students Do Their Homework? Disentangling Motivation and Study Productivity as Drivers of Human Capital Formation

Christopher S. Cotton, Brent R. Hickman, John A. List, Joseph Price, and Sutanuka Roy

ABSTRACT. Using field-experimental data (study-time tracking and randomized incentives), we identify a structural model of learning. Student effort is influenced by external costs/benefits and unobserved heterogeneity: motivation (willingness to study) and productivity (conversion rate of time into skill). We estimate academic labor-supply elasticities and skill technology. Productivity and motivation are uncorrelated. Low productivity, not low motivation, is the stronger predictor of academic struggles. School quality augments productivity and accelerates skill production. We find that dynamic skill complementarities arise mainly from children's aging and from a feedback loop between investment activity and productivity, rather than from carrying forward past skill stocks.

Keywords: Student effort; field experiment; skill technology; unobserved heterogeneity; productivity; motivation; school value added; cognitive skills; endogeneity of inputs; anchoring test scores;

JEL Codes: C93, I21, I24, J22, J24, 015

Current version: July, 2025 (first version: October 2020). **Note: previous versions of this paper circulated under the title, *"Disentangling Motivation and Study Productivity as Drivers of Human Capital Formation: Evidence from a Field Experiment and Structural Analysis."*** Gregory Sun provided outstanding research assistance and invaluable feedback with the analysis. This paper benefited from many thoughtful discussions with Martin Luccioni, Sally Sadoff, Ismael Mourifié, Barton Hamilton, Stephen Ryan, Martin Garcia Vazquez, Nicholas Buchholz, Caroline Hoxby, Petra Todd, Jeffrey Smith, Chris Taber, Joseph L Mullins, Derek Neal, Angela Duckworth, Samuel Purdy, Aloysius Siow, Felix Tintelnot, Lydia Scholle-Cotton, Daphne Hickman, Morgan Hickman, Ariadne Merchant, and Nicholas Merchant, as well as seminar participants at several seminars and conferences. The field experiment would not have been possible without the help of the anonymous school administrators and teachers at our partner schools, as well as our team of talented and enerjetic research staff, including Andrew "Rusty" Simon, Joseph Seidel, Clark Halliday, No'am Keesom, Edie Dobrez, Diana Smith, Kristen Troutman, Wendy Pitcock, Matthew Epps, Janaya Gripper, Allanah Hoefler, Justin Holz, Kristen Jones, Tova Levin, Claire Mackevicius, and and our student RAs, including Marvin Espinoza, Bonnie Fan, John Faughnan, Yuan Fei, Ian Fillmore, Greta Gol, Justin Guo, Colton Korgel, Hunter Korgel, Ethan Kudrow, Helen Li, Victor Ma, Claire Mackevicius, Janae Meaders, Mateo Portune, Denis Semisalov, Yaxi Wang, De'Andre Warren, Colleen White, and Colin Yu. We thank the editor, James Heckman, and four anonymous referees for feedback that helped improve this paper.

Christopher Cotton: Queen's University; ***Brent Hickman:*** Washington University in St. Louis; ***John List:*** University of Chicago and NBER; ***Joseph Price:*** Brigham Young University; ***Sutanuka Roy:*** Australian National University.

1. Introduction

Childhood human capital development is a three-way partnership between schools, parents, and the learner, especially during adolescence, when children assume a more proactive role in their education. There is a vast literature on factors influencing child learning, including one large branch on parental investment and/or home interventions,[1] and another large branch on the role of schools.[2] On the other hand, much less is known about how the learner's own decisions affect skill development.[3] To our knowledge, only two other papers, Todd and Wolpin (2018) and Del Boca, Flinn, Verriest, and Wiswall (2024), have structurally modeled endogenous child effort. One could argue that this missing link is uniquely crucial within the school-parent-child partnership: if the learner will not or cannot contribute, it is difficult to imagine a viable compensating factor in human capital production.

Our research focuses on a child's rational choices of *self-investment*, or how much time to devote to learning. These choices hinge crucially on learner traits that are unobservable to the researcher. For instance, suppose that *Jane*, a 6$^{\text{th}}$ grader, does not regularly complete homework and performs poorly on exams. Seeing this, common wisdom would conclude that *Jane* or her parents may lack motivation for academic pursuits.[4] If this is the issue, it may be possible to improve her outcome by providing information on the returns to schooling, or raise incentives for completing schoolwork or increasing grades. However, low motivation is not the only possible explanation for *Jane*'s low performance. Alternatively, she may be willing to work but lacks the foundational skills, ability to focus, or academic support needed to efficiently convert study time into academic improvement. In this case, raising her grades may require a sufficiently large commitment of time, so as to outweigh the rewards of higher performance. Thus, *Jane* may rationally withdraw from academics despite being no less motivated than her peers, and simply nudging beliefs or incentives may fail to address the underlying challenges holding *Jane* back.

Prior empirical work has assumed homogeneous returns to child investment, conditional on observables and prior learning. Various papers do not explicitly model optimal investment or study choices, but assume a production technology with constant coefficients on investment

[1] E.g., Cunha, Heckman, and Schennach (2010); Del Boca, Flinn, and Wiswall (2014); Fryer, Levitt, and List (2015); Chetty, Hendren, and Katz (2016); Attanasio, Cattan, Fitzsimmons, Meghir, and Rubio-Codina (2020); Attanasio, Meghir, and Nix (2020); Gayle, Golan, and Soytas (2022); Agostinelli and Wiswall (2024); and many others.

[2] E.g., Todd and Wolpin (2003); Dobbie and Fryer (2011); Cullen, Levitt, Robertson, and Sadoff (2013); Chetty, Friedman, and Rockoff (2014); Fryer (2017); Hanushek (2020); Guryan et al. (2021); Ahn, Aucejo, and James (2022); Fryer, Levitt, List, and Sadoff (2022); Luccioni (2024); and many others.

[3] An experimental literature studies responses to academic incentives (e.g., Fryer (2011), Fryer (2016) Levitt, List, and Sadoff (2016), Burgess, Metcalfe, and Sadoff (2016), and Cotton, Hickman, and Price (2022)), but generalizable insights are still sparse.

[4] Prior studies with endogenous investment have considered parental preferences (e.g., Del Boca et al. (2014); Del Boca, Flinn, and Wiswall (2016); Agostinelli, Doepke, Sorrenti, and Zilibotti (2025)) and/or teacher/child preferences (e.g., Todd and Wolpin (2018); Del Boca et al. (2024); Luccioni (2024)) as the driving determinant of investment variation, given contextual factors.

(e.g., as opposed to random coefficients, which would allow for heterogeneous returns in reduced form).[5] A smaller set of papers structurally model investment choices by parents, teachers, and/or children (see footnote 4), but assume homogeneous skill gains from an hour invested, given prior skill stock and school/neighborhood/household traits. Ours is the first study on child skill formation with structural unobserved heterogeneity (UH) in both preferences *and* in the learning returns to home study time. Multi-dimensional UH is a distinct concept from the prior literature's focus on multi-dimensional skill stocks (e.g., cognitive vs non-cognitive skills), which are partially observable through exam scores, subject to measurement error.[6] In contrast, we focus on unobservable learner traits that shape the skill accumulation process *by governing endogenous study/investment behavior*.

If skill growth depends on a child's day-to-day academic labor supply, then the challenge for policymakers, educators, and researchers is inferring whether low achievers lack unobserved motivation—willingness to spend time studying—or whether they lack unobserved productivity—conversion rate of an hour study into durable skill. Typical observational data (e.g., NLSY, PSID) include limited or no time-use information, meaning both low motivation and low productivity are plausible explanations for underachievement and cannot be separately identified. We propose a novel model of academic labor supply (Section 2) and design a natural field experiment (Section 3) to identify the structural primitives of time choice and skill development, using study-time panel data and randomized incentives.

Our framework diverges from previous time-use models by allowing for heterogeneous productivity *across* students, meaning children differ in how effectively they convert time into learning. We also allow for each successive hour of study to be heterogeneously productive *within* students, as they gain experience. Our model captures heterogeneous motivation and convex utility costs. That is, children differ by willingness to study for one hour, but they all eventually become less willing to continue as cumulative study commitment rises, due to fatigue or leisure being crowded out. Our model provides a micro-foundation for why learning may hinge on multiple latent traits.

A central implication of our academic labor-supply model is that design of compensation schemes drives learner behavior in important ways (Section 2.6). Academic incentives tend to be "piece-rate," where individuals are rewarded for outputs rather than time. If *Jane* and her classmate *Tabitha* earn top marks in math, they both see the same improvement in their job/college prospects, even if *Tabitha* required only a fraction of *Jane*'s study time. The

[5] E.g., Cunha and Heckman (2008); Cunha et al. (2010); Attanasio, Cattan, et al. (2020); Attanasio, Meghir, and Nix (2020); Agostinelli, Saharkhiz, and Wiswall (2025); Agostinelli and Wiswall (2024); Heckman and Zhou (2025).

[6] For example, Cunha et al. (2010) consider an extension of their seminal framework where a UH term, denoted as π, is correlated with observed investment. Identification under this form of investment endogeneity assumes $\pi_t \in \mathbb{R}^1$ is scalar within each period t, and their empirical application specifies it as entering the investment equation additively. Attanasio, Cattan, et al. (2020) and Attanasio, Meghir, and Nix (2020) propose instrumental variables and control functions to address omitted variable bias from UH, but require similar assumptions (scalar shocks entering log-additively). Our framework relaxes both conditions.

piece-rate structure of rewards for learning implies that latent traits interact in highly non-linear ways to produce intermediate academic labor supply and resulting skill gains. While a child's motivation determines willingness to spend an hour studying, her productivity determines *how many hours* of study are needed to reap the rewards of high achievement.

Our novel approach allows us to develop generalizable insights on student choices and academic outcomes. How responsive is academic labor-supply to external incentives, and what are the skill returns to time spent studying? How heterogeneous are adolescents in their academic productivity and willingness to substitute time from outside options toward schoolwork? What portion of this variation is attributable to contextual factors? Are these latent traits correlated? How does UH shape skill-production technology, and what role does school quality play? Empirical analysis has three main components. (*I*) Our field experiment rests on two pillars to identify and estimate labor-supply primitives (Section 4): automated tracking of web-based study allows us to quantify productivity differentials, and randomized study incentives pin down labor-supply costs and cost curvature. In a subsequent estimation stage, labor-supply parameters combined with exam and survey data, allow us to study (*II*) how productivity and motivation are influenced by external factors (e.g., outside time uses, neighborhood/household/school traits); and (*III*) how home study, latent traits, and school quality combine to shape the technology of durable skill gains.

We find evidence of vast heterogeneity in both motivation and productivity, even among high-achieving students often labeled as "gifted and talented." Much of it appears idiosyncratic to the child: 56% of productivity variation and 70% of motivation variation are unexplained after conditioning on a rich set of school/household/neighborhood covariates (Section 5). However, we can identify factors that significantly influence a child's academic traits. Notably, evidence points to higher quality schools augmenting the productivity of a child's home study time. The pooled-cohorts design of our field experiment facilitates measurement of the year-on-year change in average latent traits. Fifth-grader productivity is a third of a standard deviation lower than 6^{th}-graders, on average, while both age groups exhibit comparable average motivation. Finally, while both motivation and productivity are influential for academic labor supply, demographic groups whose exam scores have historically lagged are *not* less motivated than their higher-performing counterparts, on average.

We find that cost convexity is a significant driver of study choices: doubling a child's daily time commitment to math causes monetized utility costs to rise by a factor of more than three. Our structural model also allows us to test the hypothesis of whether adolescents' home study decisions are primarily driven by external opportunity costs of time (through crowding out of other activities) or internal psychic costs (e.g., burnout, aversion to work, etc., Section 5.1.1). Our data favor the latter as dominant, and are consistent with a model of holistic time choice where adolescents are not heavily time-constrained at home.

Turning to skill production technology (part (*III*) above, Section 6), our specification departs from standard parametric assumptions by previous studies, and adopts a more flexible functional form instead. We are able to do so for three reasons. First, we can control for UH, pre-estimated at the student level within the labor-supply model. Second, UH estimates have economically interpretable scales, which simplifies measurement-error issues that arise when exam scores are used to measure skills. Third, our web-based tracking of learning activity eliminates a key source of measurement error in investment data that has plagued previous work. Moreover, our website separates investment into two components: time spent and learning assignments completed. This allows us to study the question of whether skill growth arises through simply spending more time studying math at home, or through completing a larger volume of homework and solving more practice problems. The evidence strongly favors the latter, which underscores the policy relevance of productivity differentials: it is not merely one's study time that leads to learning, but rather, the *intermediate inputs produced by one's time* are primarily responsible for augmenting math skill.

To avoid the hard-to-evaluate assumptions embedded in standard functional forms (e.g., Cobb-Douglas, CES (constant elasticity of substitution), and trans-log), we specify skill production technology as a complete polynomial in a two-dimensional investment vector (time spent and assignments completed), where slope and intercept terms are functions of prior skill stock, UH, and student covariates. Thus, we can capture novel interactions including (*i*) between different components of UH in determining endogenous investment; (*ii*) between UH and investment; and (*iii*) between UH and external factors such as school quality. For example, we find that students with higher productivity not only choose to complete more learning assignments at home, but they also convert a fixed volume of work into larger durable skill gains. Since we do not impose *a priori* monotonicity restrictions on the inputs of the production technology, we find that the partial derivative of time is actually negative, even though its total derivative is positive. In other words, while study time is overall productive of useful intermediate inputs (practice problems solved), all else equal (including completed work), it is a moderating factor for study efficiency.

Our work also speaks to key issues within the item-response theory (IRT) literature (see Cunha, Nielsen, and Williams (2021)). We identify UH by combining panel data on work times with revealed-preference methods and variation in study incentives. This addresses a key problem from IRT that the mapping between exam scores and the true cardinal scale of latent traits (which they attempt to measure) is unknown. Following Cunha and Heckman (2008), various empirical analyses have normalized the unknown cardinal scales of latent skill (known as *anchoring*) by assuming a linear relationship exists between skill stocks/exam scores and interpretable outcomes (e.g., adult earnings). In contrast, we anchor the cardinal scale of UH to economically interpretable variables (time and earnings per unit of time) using a micro-founded model of optimal home-study choices, which determine exam scores. Skill technology estimates indicate decreasing returns to scale, where math score gains of a fixed margin become more labor-intensive as one's baseline score rises.

Our combined results point to eight distinct mechanisms affecting the rate of skill growth as a child develops (summarized in Section 6.3.4). These imply a nuanced interpretation of dynamic skill complementarity, or the notion that "skills beget skills" (Cunha and Heckman (2007)). We find that endogenous investment activity is self-reinforcing by improving one's productivity over time, which spurs further investment and accelerates skill growth in various ways. Carrying forward higher skill stocks from previous learning *decelerates* skill growth somewhat, though not by enough to overcome the other seven advantageous mechanisms.

Finally, we contribute to the literature on school value-added (VA) (e.g., Koedel, Mihaly, and Rockoff (2015); Abdulkadiroglu, Pathak, Schellenberg, and Walters (2020)). Our approach has several advantages. First, we can directly control for selection on unobservables when estimating school VA. Second, school VA (and IRT) studies have assumed scalar latent ability (see Koedel et al. (2015) and Cunha et al. (2021)), whereas we allow for multiple latent traits with distinct and interacting contributions to skill growth. Third, our approach allows us to explore specific mechanisms through which school VA may operate: (*i*) by augmenting study productivity; (*ii*) by augmenting motivation; (*iii*) by directly augmenting skill production independently of home study; and (*iv*) by increasing the rate at which homework accomplishment translates into durable skill. We find that three of these four channels ((*i*), (*iii*), and (*iv*)) play a meaningful role. Our estimates point to substantial heterogeneity of school VA treatment effects through channels (*i*) and (*iv*).

2. A Theoretical Framework for Academic Labor Supply

We propose a quantitative framework of rational learner choice in the spirit of influential qualitative models from psychology on the mechanics of learning.[7] That literature frames skill acquisition as a sequence of learning-by-doing tasks: a math student attends class, is assigned homework, chooses how much of each assignment to complete, and iterates the daily process until the end of the course. Her skill gain is then determined by the cumulative volume of interim learning tasks she completed over time. Our goal is to formally quantify a child's learning process, which we separate into two main components: academic labor-supply (Sections 2–5) and skill production technology (Section 6).

2.1. Time Requirements of Learning Task Completion.

For each child i, let $A_i \in \mathbb{N}$ denote the number of learning tasks that i completes within a fixed period of time. "Learning tasks" are discrete packets of work with comparable content (e.g., middle-school math) and size. The total volume of within-period learning task completion A_i is central to the model, while the precise partitioning of tasks is not. A task can represent a single math problem, a block of problems, or a larger assignment. The notion of "completing" a learning task involves some form of quality control: students do not receive credit simply for supplying answers to math problems, but for the quantity they solved correctly. Similarly, the precise

[7]E.g., "mastery theory" (Carroll (1963)), "expectancy-value theory" (Eccles et al. (1983)), and "achievement mechanics" (Duckworth, Eichstaedt, and Ungar (2007)).

partitioning of a "time period" is not crucial—e.g., a week, a month, or a semester—provided it is short enough that a child's baseline decision-relevant characteristics can be thought of as relatively stable within a period. Our model focuses on trade-offs of allocating different portions of a finite, *within-period time endowment*, $\overline{T}$, to math study.

Definition 2.1. Student i's *academic productivity type*, $\theta_{pi} > 0$, is inversely proportional to the (within-period) time rate at which i is able to complete learning tasks.

Although rate of progress is inversely proportional to θ_{pi}, for simplicity we just refer to it as one's "productivity." Learning task completion happens within the unstructured home environment—in the absence of time constraints, with access to help from parents, tutors, textbooks, and other feedback—at a rate determined by θ_{pi}. This activity builds durable skill κ_i, measured by exam scores S_i derived from a controlled, time-constrained testing environment. Thus, θ_p and κ are related, but are not the same thing: productivity is an unobservable trait impacting endogenous investment, which in turn drives skill growth. Skill stock κ_i is measured by S_i, which *constrains one's time* while measuring unassisted *accuracy* in problem solving. In contrast, productivity applies to the unstructured environment at home, where accuracy is constrained—students get credit only for practice problems correctly solved—and θ_{pi} determines *time required* to achieve a *fixed standard*.

It will be useful to define a chronological index, $a = 1, 2, \ldots, A_i$, for child i's completed learning tasks. We model time required for the a^{th} completed learning task as a random variable whose data-generating process is described as follows:

$$\tau_i(a) \equiv \mathcal{T}_0 \times \mathcal{T}_1^{\mathbb{1}(a=1)} \times \theta_{pi} \times a^{-\varphi} \times U_{ia}, \quad \mathcal{T}_0,\ \mathcal{T}_1,\ U_{ia} > 0, \quad a = 1, \ldots A_i, \quad i = 1, \ldots, N. \tag{1}$$

Here, $\mathcal{T}_0$ determines mean baseline completion time, $\mathcal{T}_1$ is a startup cost on the first task, and the fixed effect θ_{pi} scales median completion time up or down. The transitory shock, U_{ia}, is *iid* across tasks and represents unpredictable fluctuations in difficulty, distractions, etc. This is a realistic representation of the data—within-student completion times vary substantially across tasks (Table 1)—and reflects common academic experience: sometimes a dreaded math problem yields a quick solution, and other times a learner's initial optimism gives way as an exercise drags on.[8] Upper-case U denotes random outcomes and lower-case u denotes specific realizations. Subscripted "F" denotes exogenous distributions, while subscripted "G" is for endogenous distributions. We assume the shock distribution is well-behaved:

[8]We assume a represents chronology of completed learning tasks from among a menu of comparable options, rather than a series of developmental benchmarks, as in Heckman and Zhou (2025). In our setting, the final volume of work A_i is central to skill growth, while the identities of the individual homework problems $a = 1, \ldots, A_i$ are not. For example, suppose a math course features 20 assignments, each with 5 practice problems indexed (i)-(v). On assignment 1, *Tabitha* completes all in the order $\{(ii),(iii),(i),(v),(iv)\}$, while *Jane* completes three in the order $\{(i),(ii),(iii)\}$, but skips problems (iv) and (v). Thus, chronological indices $a \in \{1,2,3\}$ for *Tabitha* and *Jane* map to different problems of homework 1, and $a \in \{4,5\}$ for *Jane* will map to problems on homework 2 or later. We assume learning tasks in the assigned list are ex-ante identical by expected work times, while allowing for ex-post differences in time requirements.

Assumption 1. *Time shocks U_{ia} follow heteroskedastic distribution $F_u(u_{ia}|\theta_{pi})$ with continuous density f_u that is bounded away from zero on nonempty support $[\underline{u}, \overline{u}] \subset \mathbb{R}_+$.*

The term $a^{-\varphi}$ in (1)—commonly known as a *learning curve* or "Wright's Law"—allows for acceleration ($\varphi > 0$) or decay ($\varphi < 0$) in one's rate of progress through homework, as the number of previously completed assignments grows.[9] implies an interim law of motion: one's effective productivity for the a^{th} unit of work is $\rho_i(a) \equiv \theta_{pi} \times a^{-\varphi}$, where θ_{pi} is the idiosyncratic productivity baseline, and $a^{-\varphi}$ is the systematic component.

2.2. **Utility Costs of Time.** In what follows, it will be useful to define $T_i(a) \equiv \sum_{a'=1}^{a} \tau_i(a')$ as a random variable that aggregates all work time through the a^{th} completed learning task, with T_i as shorthand for $T_i(A_i)$. We model costs of diverting some fraction of within-period time endowment $\overline{T}$ toward math and away from the most profitable outside use.

Definition 2.2. Student i's *academic motivation type* $\theta_{mi} > 0$ is proportional to her utility cost of supplying t hours to study, given by $C_i(t; \theta_{mi}) \equiv \theta_{mi} c(t)$, $t \leq \overline{T}$.

Although willingness to spend time studying is inversely related to θ_{mi}, for ease of discussion we refer to it simply as "motivation." Multiplicative separability, one of the most common utility forms in empirical principal-agent studies,[10] implies that i's labor-supply cost levels and marginal costs are increasing in θ_{mi} for any $t \geq 0$. It provides a significant degree of empirical tractability, while allowing the researcher to still be quite flexible on specifying the functional form for the common component of costs, $c(t)$.[11] Assumption 2 establishes regularity conditions that ensure well-behaved leisure-study decisions.

Assumption 2. *Costs are twice differentiable, $c \in \mathcal{C}^2$; increasing, $c'(t) > 0$; convex, $c''(t) > 0$; with unbounded marginal costs, $\lim_{t \to \overline{T}} c'(t) = \infty$; and normalizations $c(0) = 0$ and $c'(0) = 1$.*

Cost convexity has a simple interpretation: each successive study hour is more costly than the previous one, because marginal utility of outside options rises (as consumption is crowded out), *and/or* because the direct marginal psychic cost of math effort rises (burnout) with the fraction of $\overline{T}$ devoted to study. Unbounded marginal costs ensures finite study choices, but is also intuitive: if a period is a week, then 168 hours of continuous math study would require extreme levels of physical strain and sleep deprivation. Our cost specification does not derive

[9]Wright (1936) pioneered the study of learning-by-doing (see Thompson (2010) for a survey). Equation (1) follows the main specifications in Hirsch (1956), Arrow (1962), and Klenow (1998), with additional startup costs and production shocks.

[10]E.g., in labor supply (Saez, 2002; Chetty, 2012; Kleven & Waseem, 2013; D'Haultfoeuille & Février, 2020; Hedblom, Hickman, & List, 2022); nonlinear pricing (Luo, Perrigne, & Vuong, 2018; Attanasio & Pastorino, 2020; Brugués', 2024; Bodoh-Creed, Hickman, List, Muir, & Sun, 2024); and firm regulation (Kang & Silveira, 2021; Chen, Jia, Martin, & Silveira, 2024).

[11]Todd and Wolpin (2018) assume a utility function that is observationally equivalent to our specification for the special case where $c(t) = t^2$. Bodoh-Creed et al. (2024) show evidence within a similar principal-agent model that additively separable utility and multiplicatively separable utility produce UH estimates from exogenous price shifts (analogous to inferring θ_{mi} from observed study choices) that are very similar. Sun (2024, Theorem 3) formalizes this result.

from an explicit model of a child's allocation of time across *all* activities (though we consider our estimates in the context of such a model in Section 5.1.1). There are two advantages of this approach. First, it provides a simple and tractable model of how a child trades off external incentives to study math. Second, it allows for some generality, as our specification of the cost function need not pre-suppose a specific model of all time allocation choices. Accordingly, we interpret the parameter θ_{mi} as a reduced-form index which subsumes child i's idiosyncratic opportunity costs, as well as her direct psychic costs.

Similarly as how marginal productivity $\rho_i(a)$ evolves with interim investment, the marginal cost of effort at the beginning of task a evolves as $\psi_i(a) \equiv \theta_{mi}c'(T_i(a-1))$. We assume idiosyncratic baseline motivation θ_{mi} is fixed and stable over the short run, as this delivers computational tractability. This assumption is in keeping with stylized conventions from the discrete-time dynamic investment literature (footnotes 4 and 5), though our implementation is potentially less restrictive given the shorter span of a period in weeks rather than years. However, convexity implies $\psi_i(a)$ is an interim law of motion for marginal labor-supply cost.

2.3. **Payoffs and Optimal Choices.** A payoff function Π_i encompasses external "carrots" and "sticks" that entice child i to study. Parents may inculcate intrinsic value of achievement, or offer tangible rewards (punishments) for (non-)completion of work. Classmates wield peer influence, and schools motivate students with grades for homework completion and pre-announced exams. Businesses, organizations, and universities offer achievement-based admissions, scholarships, internships, or job prospects that improve expected flows of future monetary and psychic income from a desirable career (Becker (1993)).

Assumption 3. *An increasing piece-rate payoff function* $\Pi_i(a)$ *governs external incentives, such that* $\exists\, \overline{\pi} < \infty$ *where* $\Pi_i(a) - \Pi_i(a-1) \leq \overline{\pi}$, $\forall a \geq 2$.

In choosing (A_i, T_i), child i solves an optimal stopping problem (Wald (1945) and Arrow, Blackwell, and Girshick (1949)) by recursively comparing costs and expected benefits of an additional completed assignment. The chronological index a plays three payoff-relevant roles by controlling interim evolution of marginal productivity $\rho_i(a)$, marginal costs $\psi_i(a)$, and marginal payoffs $(\Pi_i(a)–\Pi_i(a–1))$. Given some history of $(a–1)$ completed assignments, a student optimally determines the maximum time t^*_{ia} that she is willing to spend on the a^{th} task. Equation (1) implies that success probability on task a, given t units of time input, is

$$F_s(t; a, \theta_{pi}) \equiv \Pr\big[\tau_i(a) \leq t\big] = \Pr\left[U_{ia} \leq \frac{t}{\left(\mathcal{T}_0 \mathcal{T}_1^{\mathbb{1}(a=1)} \rho_i(a)\right)}\right] = F_u\left(\frac{t}{\left(\mathcal{T}_0 \mathcal{T}_1^{\mathbb{1}(a=1)} \rho_i(a)\right)}\bigg|\theta_{pi}\right), \tag{2}$$

with first derivative $f_s(t;a,\theta_{pi})$. A learner's decision problem is defined by the Bellman equation,

$$\mathcal{V}\big(a-1,T_i(a-1)\big) = \max_{t\geq 0}\Bigg\{\int_0^t\Bigg[\Big(\Pi_i(a)-\Pi_i(a-1)+\mathcal{V}\big(a,T_i(a-1)+\tau\big)\Big) \\ -\theta_{mi}\Big[c\big(T_i(a-1)+\tau\big)-c\big(T_i(a-1)\big)\Big]\Bigg]f_s(\tau;a,\theta_{pi})d\tau \\ -\theta_{mi}\Big[c\big(T_i(a-1)+t\big)-c\big(T_i(a-1)\big)\Big]\Big[1-F_s(t;a,\theta_{pi})\Big]\Bigg\}. \tag{3}$$

The first term inside the integral is payoffs from work, including incremental rewards from the current task, and retention of continuation value for future work if i completes the a^{th} task before exhausting her willingness to continue. The integral and density term are to account for the fact that i may complete task a before exhausting her planned maximum work time t^*_{ia}. Finally, costs are split between the integral and the final term in the curly brackets because there is also positive probability that i may not complete task a before exhausting t^*_{ia}. Optimal choice is implicitly defined by the first-order condition[12]

$$f_s(t^*_{ia};a,\theta_{pi})\Big(\Pi_i(a)-\Pi_i(a-1)+\mathcal{V}\big(a,T_i(a-1)+t^*_{ia}\big)\Big) = \theta_{mi}c'\left(T_i(a-1)+t^*_{ia}\right). \tag{4}$$

If the learner completes task a at time $t<t^*_{ia}$, then she pauses and re-optimizes Bellman equation (3) with $(a,T_i(a-1)+t)$ as the updated state variables. Otherwise, if devoting time $t<t^*_{ia}$ does not complete assignment a, she continues to work and her stopping time t^*_{ia} balances the expected marginal benefit of continuing (including expected future profit opportunities) against the deterministic marginal cost. If she reaches t^*_{ia} before her a^{th} success, she discontinues work for rest of the period, reserving her remaining endowment $\overline{T}-T_i(a-1)-t^*_{ia}$ for non-math activities, and her final work volume is $A_i=(a-1)$.[13]

2.4. **Durable Skill Production.** Although the optimal stopping problem does not yield a closed-form solution, it will be useful to represent resulting choices using labor-supply curves, $A_i=L_A(\theta_{pi},\theta_{mi},\Pi_i,\boldsymbol{u}_i)$ and $T_i=L_T(\theta_{pi},\theta_{mi},\Pi_i,\boldsymbol{u}_i)$, based on UH, incentives, and a shock sequence $\boldsymbol{u}_i=\{u_{ia}\}_{a=1}^{\infty}$. We also denote total investment activity by $\boldsymbol{I}(\boldsymbol{\theta}_i,\Pi_i,\boldsymbol{u}_i)\equiv[A_i,T_i]^\top$, where $\boldsymbol{\theta}_i=[\theta_{pi},\theta_{mi}]^\top$. At the end of the period, i's *skill growth* is defined as

$$\Delta\kappa_i\equiv\kappa_i^{post}-\kappa_i^{pre}=\mathcal{F}[\kappa_i^{pre},\boldsymbol{I}(\boldsymbol{\theta}_i,\Pi_i,\boldsymbol{u}_i),\boldsymbol{\theta}_i,\boldsymbol{X}_i,\varepsilon_i]\,, \tag{5}$$

where κ_i^{post} is incremented post-investment skill, κ_i^{pre} is pre-existing skill, $\boldsymbol{X}_i$ denotes contextual factors, and and ε_i is a skill growth shock. There are two primary innovations in

[12]If the left-hand side of (4) is less than the right-hand side given $a=0$ and $T_i=0$, then a corner solution obtains.

[13]We abstract from directly modeling non-terminal units of work that are attempted prior to A_i but never completed, as these events are empirically rare (see Section 3.2.2). Thus, t^*_{ia} is maximal time i can rationally spend in pursuit of an a^{th} completion.

our skill technology framework. First, and most importantly, we allow for endogenous investment $\boldsymbol{I}$ to depend on a 2-dimensional UH vector $\boldsymbol{\theta}_i$, and our explicit model of optimal investment provides a rationale for why this can account for meaningful phenomena ruled out in previous models (see Sections 2.5.1 and 2.6 below). Second, we allow for the UH vector $\boldsymbol{\theta}_i$ to also directly enter the production function $\mathcal{F}$, aside from its influence on academic labor supply.[14] We defer further discussion and empirical analysis of $\mathcal{F}(\cdot)$ to Section 6.

2.5. **Model Discussion.** Unconditional on shocks, there is a non-degenerate distribution of optimal choices, $G_{AT}(A_i, T_i|\theta_{pi}, \theta_{mi}, \Pi_i)$, given some $(\theta_{pi}, \theta_{mi}, \Pi_i)$ triple. Two key model implications are relevant to identification. First, if a child proceeds to the $(a+1)^{\text{th}}$ task, her baseline cost rises since $T_i(a) > T_i(a-1)$ with probability one. Cost convexity implies rising marginal costs $\psi_i(a) > \psi_i(a-1)$, and bounded marginal payoffs imply that willingness to work eventually declines: $\forall i\ \exists a_i^*$ such that $t^*_{i,a+1} < t^*_{ia}\ \forall a \geq a_i^*$. Second, the model predicts labor supply is decreasing in θ_{mi}, conditional on shocks $\boldsymbol{u}_i$, productivity θ_{pi}, and incentives Π_i. This follows from Milgrom and Shannon (1994, Theorem 4) and because our cost specification satisfies the Milgrom-Shannon single-crossing condition, implying that the marginal incentive to work is decreasing in θ_{mi}. By similar logic, labor supply is also decreasing in θ_{pi}, conditional on shocks $\boldsymbol{u}_i$, motivation θ_{mi}, and incentives Π_i.

Our decision model is a non-stationary dynamic program because of cost convexity and history-dependence of the state variable $T_i(a-1)$. Continuation value $\mathcal{V}\big(a, T_i(a-1) + \tau\big)$ updates continuously with time on task a, and the distribution of payoffs on future work $a' > a$ is not known until the student finally completes a. The model is computationally taxing for the researcher, but requires only simplistic thinking on the part of the learner. A child need only comprehend at each point in time her marginal incentives to complete task a (and potentially more work afterward), the variability of completion times for tasks $a' \geq a$, and how taxed she feels from her cumulative study time through the present.[15]

Finally, our choice to model payoffs $\Pi_i(A_i)$ as a function of completed assignments is consistent with previous literature that has modeled payoffs as arising from skill stocks κ_i (footnote 4), because A_i is an intermediate input that drives skill growth. Under the assumption that no learning can happen without the learner's participation, this framing of the payoff function is general enough to admit various interpretations, including payoffs arising from strategic models (e.g., Todd and Wolpin (2018); Del Boca et al. (2024)) or

[14] Consequently, our focus on evolving cognitive skill stock, κ_i, is not necessarily restrictive, relative to the seminal framework of Cunha et al. (2010), where cognitive and non-cognitive skill stocks—both functions of scalar UH—influence the returns to investment in cognitive skill. One interpretation of the third argument of $\mathcal{F}$ is that it may account for the role of non-cognitive skill stock on cognitive (math) skill evolution. Moreover, the third argument of $\mathcal{F}$ is a generalization on Cunha et al. (2010), as it allows for direct influence of multiple latent traits on cognitive skill evolution, conditional on current cognitive skill stock.

[15] A simpler one-shot decision model would require stronger assumptions: children must rigidly stick to an ex-ante study plan, A_i, regardless of whether unexpectedly (dis-)advantageous time shocks are encountered along the way.

single-agent models (e.g., Del Boca et al. (2014); Agostinelli, Doepke, et al. (2025)), but it does not require us to assume a specific mechanism through which study affects utility.

2.5.1. *Comparison to Existing Frameworks.* Our labor supply model (equation (3)) and skill technology framework (equation (5)) provide a micro-foundation for why multiple unobserved traits $(\theta_{pi}, \theta_{mi})$ may play a role in skill growth $\Delta\kappa_i$, and why their separate contributions may interact in non-linear ways. In that sense, our model generalizes relative to prior literature, which can be partitioned into four categories: (*I*) studies that do not model investment choice and assume exogenous shocks to skill growth (i.e., they assume no UH); (*II*) studies that do not model investment choice but allow for endogeneity due to 1-dimensional UH; (*III*) studies that model investment choice but assume exogenous shocks to skill growth (i.e., they assume no UH); and (*IV*) studies that model investment choice and allow for 1-dimensional UH on the part of learners.[16] Frameworks in categories (*I*) and (*III*) omit the third argument in equation (5) and restrict UH to be a function of observables $\boldsymbol{\theta}_i = \mathcal{H}(\boldsymbol{X}_i)$. Category (*II*) and category (*IV*) models assume $\boldsymbol{\theta}_i \in \mathbb{R}^1$ and restrict or omit the role of the third argument in equation (5) (See footnote 14). Our 2-dimensional UH model implies a non-degenerate distribution of counterfactual investment responses under a change of incentives, conditional on prior skill stock and observables, while the no-UH or scalar-UH models assume that labor-supply responses to incentive changes are conditionally deterministic.[17]

Relative to equation (5), the dimension that has been absent from previous empirical work on skill production is unobserved productivity differences.[18] Categories (*I*)–(*III*) assume constant coefficients on investment as opposed to a random-coefficients specification, for example, which would capture heterogeneous returns. Category (*IV*) studies interpret the UH model components as preferences—tastes for achievement (Todd and Wolpin (2018)) or child discount factors(Del Boca et al. (2024))—but assume constant returns to time invested, conditional on a child's classroom or parents. This begs the question, is the additional degree of freedom in our model needed to rationalize education data? In related work, Cotton, Hickman, List, and Sun (2025) propose a formal test, and find that a majority of productivity variation is unexplained after accounting for traditional observables such as previous learning, age, classroom traits, neighborhood traits, and household traits.

[16] Category (*I*) papers include Cunha and Heckman (2008); Agostinelli and Wiswall (2024); Heckman and Zhou (2025); Agostinelli, Saharkhiz, and Wiswall (2025); category (*II*) includes Cunha et al. (2010); Attanasio, Cattan, et al. (2020); Attanasio, Meghir, and Nix (2020); category (*III*) includes Del Boca et al. (2014, 2016); Agostinelli, Doepke, et al. (2025); and category (*IV*) includes Todd and Wolpin (2018); Del Boca et al. (2024).

[17] As articulated by Todd and Wolpin (2018), in models built on the Ben-Porath (1967) framework, "*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.*" In contrast, our framework allows many unobserved $(\theta_{pi}, \theta_{mi})$ pairs to produce the same skill stock.

[18] Of course, studies in categories (*I*)–(*IV*) allow for heterogeneous investment returns based on observables such as prior test scores, age, and home/school/neighborhood traits.

2.6. **Illustrative Examples: Incentives and Study Choice.** Piece-rate incentives dominate academics: if *Jane* and *Tabitha* each complete all homework assignments and receive top marks in math, no distinction is made if *Jane* required many times *Tabitha*'s home-study commitment. Both receive the same improved prospects for a desired college seat, job, etc. With piece-rate decisions, learning depends both on a child's willingness to study an hour (θ_{mi}), and also on *how productive that hour will be* (θ_{pi}) for reaping output-contingent rewards. Our model calls into question common conceptions of observed student behaviors. Now suppose *Jane* completes only 50% of her math assignments while *Tabitha* completes 100%. All else observably equal, prevailing wisdom (see footnote 4) interprets *Jane*'s choices as signaling low study motivation. Our model admits this as a plausible explanation, but it is also plausible that *Jane* completes less homework despite being *more* motivated ($\theta_{m,Tabitha} \geq \theta_{m,Jane}$), if she is sufficiently less productive ($\theta_{p,Tabitha} < \theta_{p,Jane}$). Both explanations are observationally equivalent, given academic records for *Tabitha* and *Jane*.

The model also suggests a re-thinking of the term "effort." In the example above many would say *Tabitha* put forth more "effort," based on her larger observed volume of work. However, if *Jane* is less than half as productive, then she actually spent *more* time working, and in that sense exerted greater "effort" than *Tabitha*. Moreover, even if *Tabitha* and *Jane* study the same amount of time they may incur unobservably different costs. If *Tabitha* is more averse to math (higher direct costs), or has less free time due to family duties like chores or sibling care (higher opportunity costs), then each hour of her study may be more taxing than for *Jane*. Thus, multiple dimensions of UH implies that no one-to-one mapping exists between grades/test scores (or even time surveys) and unobserved measures of true effort, $\theta_{mi} c(T_i(\theta_{pi}, \theta_{mi}, \Pi_i, \boldsymbol{u}_i))$, which reflect time spent *and* idiosyncratic costs of time.

3. A Field Experiment to Identify Structural Academic Labor Supply

3.1. **A Structurally-Motivated Field Experiment.** Our study contributes to a nascent literature using field experiments to estimate structural UH.[19] Our experiment was *not* designed to test the treatment effect of paying students to study (e.g., Levitt et al. (2016)). Rather, it is designed to generate the data needed to estimate a model of academic labor supply with UH. The first component of our identification strategy is to obtain time data across learning tasks, enabling use of panel-data methods to pin down productivity. Second, we use randomized incentive variation to identify preferences and cost curvature.

3.2. **Experimental Design Details.** We developed a web-based learning platform with age-appropriate math content designed by pedagogy experts (see Online Appendix B.2). Students could complete up to 80 math learning tasks, and they had access to the website for 10 days (2 school weeks). Randomized piece-rate incentives were offered, based on the

[19]E.g., Augenblick, Niederle, and Sprenger (2015), Rao (2019), DellaVigna, List, Malmendier, and Rao (2022), Hedblom et al. (2022), Bodoh-Creed et al. (2024)

number of activities completed successfully; students could choose to do as much or as little as they wished. Our website monitored user activity and tracked completions.

School administrators and teachers cooperated with the research team during AY2013-2014, administering in-class exams/surveys and distributing letters with personalized website login credentials and incentives. A small fraction of students opted out ($< 5\%$),[20] but teachers and parents were generally enthusiastic about our study as a learning opportunity. We explain the basics of experimental design and data collection here; additional details are in Online Appendix B. Data analyses focus solely on children in non-special needs classes.

3.2.1. *Study Sample.* A total of 1,676 5[th]- and 6[th]-grade students participated, and both age cohorts were pooled together for all study aspects, including website, surveys, and assessments. Pooling served multiple purposes. First,it gave us a larger bank of learning materials to draw from. Second, it provided a wide swathe of challenging and basic math content for studying a diverse population. Third, it facilitated a comparison between age groups, allowing us to cleanly estimate the effect of an additional year of schooling on skill formation. Online Appendix B.2 presents a side-by-side harmonization of grade-specific math topics.[21]

We worked with 3 public school districts in the suburbs of the Chicago metropolitain statistical area: *District 1* (46% of the sample), *District 2* (27%), and *District 3* (27%). We summarize school traits for context in Table OS.4 (Online Appendix). District 1 was above-average on financial resources per student, faculty compensation, teacher qualifications, fraction of budget spent on instruction, and student outcomes. District 2 was close to the state averages, and District 3 lagged considerably behind on various resource dimensions. There is pronounced residential sorting by socioeconomics and race as in many other U.S. population centers. District 1 households are quite affluent by income and wealth, and mostly White and Asian. District 3 serves a low-income, low-wealth population that is mostly Black and Hispanic. District 2 households are largely middle class, and mixed race.

3.2.2. *Website Structure.* Students accessed our website through personalized login credentials, enabling us to track individual usage. Our mobile-friendly website accommodated various connection modalities.[22] It provided 80 learning tasks, each with 6 multiple-choice questions from our bank of 672 unique math problems (see Online Appendix B.2). Task size was chosen based on feedback from adolescent pilot-study subjects. Passing a task required answering at least 5 of the 6 problems correctly. Each student was allowed unlimited attempts at a given task, but for each new attempt, the orders of the questions and answer choices were randomly perturbed. Pilot participants reported that this made attempts at

[20]Participation was on an opt-out basis: students in each math class were included in the study unless the child or a parent declined. See Online Appendix B.1 for additional discussion on ethical vetting for our field experimental procedures.

[21]Concerns about pooling age cohorts are mitigated by striking curriculum similarities between grades 5 and 6. Our study attracted enough participation from all ages for adequate statistical power (see Sections 3.4, 4.4, and 6).

[22]Page-load frequencies were roughly 3/4 from desktops/laptops, 1/5 from tablets, and 1/20 from smartphones.

gaming the system (e.g., repeated guessing) unprofitable. Problems were selected at random from our question bank, so that relative difficulty was impossible to predict from one task to the next. Only one math problem appears per page, so we get a high-frequency log of work times for each child.[23] Thus, we obtain data on total website time T_i, task completions A_i, and a within-child panel of task-specific work times $\{\tau_{ia}\}_{a=1}^{A_i}$.

Although 7% of students reported not having a regular home internet connection, they were not statistically less likely to participate on the website, suggesting they had other ways of accessing the internet.[24] Having a regular home internet connection was not a significant predictor of website task completion, after controlling for school district, socioeconomic status, regular homework time, math attitude, and incentives (see Online Appendix B.4). These results suggest that differences in internet access were not a major concern.

3.2.3. *Incentives and Randomization.* Incentives took the form $\Pi_i(A_i) = (\pi_0^i + \pi_1^i A_i)\mathbb{1}(A_i \geq 2)$; to ensure a within-student panel, participants were informed they must complete at least 2 learning tasks to receive any payment. Each child was individually randomized to one of three contracts—$(\pi_{01}, \pi_{11}) = (\$15, \$0.75)$, $(\pi_{02}, \pi_{12}) = (\$10, \$1.00)$, and $(\pi_{03}, \pi_{13}) = (\$5, \$1.25)$—thus ensuring treatment variation within schools, grades, and classrooms.[25] Moving forward, superscripts denote i's contract assignment, while subscripts denote the fixed payment parameters for each contract; i.e., $(\pi_0^i, \pi_1^i) = (\pi_{0j}, \pi_{1j})$ indicates that i operated under contract-j incentives. The website prominently displayed for users (on the main menu page) their piece rates, as well as total accrued earnings and remaining potential earnings. Adequate subject motivation depends on the ratio of payoffs to work for each learning task. Partitioning each learning task into small units (6 problems each) encouraged participation and facilitated precise panel-data inference on θ_{pi}. Proportional variation in offers was significant: contracts 2 and 3 had piece rates 33% and 66.7% higher, respectively, relative to contract 1.

Our block randomization separated students into race-gender-school-grade bins. Within each, we stratified by pre-test scores and randomly assigned consecutive blocks of 3 students to contract groups 1, 2, or 3, individually. The algorithm repeated this process 50,000 times, and selected the candidate assignment that minimized correlations between treatment and balance variables (see Online Appendix E for balance table).

3.2.4. *Classroom Tests and Surveys.* Teachers proctored in-class, standardized math assessments before and after the sample period. Exams comprised 36 questions, chosen to balance 5^{th}- and 6^{th}-grade content, sub-topics, and difficulty level, using the same sources as website materials (see Appendix B.2). All students were given 35 minutes to complete as many problems as possible. Students also completed surveys on myriad individual factors, including

[23]Online Appendix B.5 provides additional detail on website time measurement, including treatment of edge cases.

[24]Among this group, computer-based pageloads were 14% lower and smartphone pageloads were higher by a similar margin. The 95% CI for participation rate among them, $[0.32, 0.50]$, contains the participation rate 0.447 for the full sample.

[25]Base payments varied inversely with marginal wage to mitigate concerns of fairness by participants.

attitudes, extracurricular activities, typical study time, availability of homework support, and more. We also gathered socioeconomic indicators from the American Community Survey for each of the roughly 160 US Census block groups where students lived. Since exams and student survey data play no role in labor-supply identification, we defer a more complete discussion of these data to Sections 5, and 6, and Appendix B.3. There we study external influences on student traits (θ_p, θ_m) (Section 5), and estimate skill technology (Section 6).

3.2.5. *Experiment Timeline.* The experiment took place as follows. (1) Teachers disseminated parental information sheets and assent forms two weeks prior to the pre-test. (2) Students received their own assent form, and took an in-class pre-test and survey administered by their teachers. (3) Students were randomly assigned incentives and provided with information about the experiment, website, and their earnings potential. (4) For a 10-day period, students had access to the website to complete learning tasks, success on which was compensated according to their assigned offer. (5) Teachers administered a post-test and survey in class. (6) Payments were mailed out within two weeks of the post-test.

3.3. **Caveats and Challenges.** We do not observe student time use and academic activity outside of the experiment. Partially observed learning/investment activity is a ubiquitous problem encountered by many previous studies (see footnotes 4 and 5). Our field experimental design solves this issue by assigning many children the same known external incentive variation. Structural identification (see Section 4.2) requires only that incentives are independent of a child's teacher, school, and home, which randomization ensures. The other key element is that experimental math activities are comparable to formal coursework, which is true by design (see Appendix B.2). Provided these two criteria are met, concurrent coursework outside of the experiment merely changes the interpretation of the motivation parameter, somewhat. In a hypothetical, "ideal" experiment, where the researcher controls all incentives and observes all math activity, a child's willingness to allocate time is judged relative to the baseline of *zero math activity.*[26] In our field experiment, θ_{mi} represents marginal willingness to allocate time to *extra math* beyond regular schoolwork, and θ_{mi} subsumes i's baseline marginal cost. Therefore, structural estimates remain informative for policy analyses focused on improvements *relative to the status quo*, which requires increased study.

Despite this caveat, rich data allow us to look beyond the basic extracurricular interpretation of experimentally inferred θ_m. UH terms $(\theta_{pi}, \theta_{mi})$ are fixed effects encompassing all factors relevant to i's productivity and motivation that are stable over the short-run. In Section 5 we project UH estimates onto student observables to study how they are impacted by demographics, home/neighborhood traits, and a dozen outside time-use controls including formal coursework commitments (see Section 5.1.1). In principle, the researcher could thus remove spurious apparent motivation by computing residual $\hat{\theta}_{mi}$, net of time constraints.

[26]The scale normalization $c'(0) = 1$ (Assumption 2) implies that $\psi_i(1) = \theta_{mi}$, so one's motivation type is directly interpretable as baseline marginal cost at zero website math activity.

Another concern is whether students responded to our incentives by neglecting regular schoolwork. Our educator partners reported no perceptible reduction in homework completion rates during the sample period, and we present survey evidence in the next section consistent with that. Finally, there is a question of extensive-margin choice: holding experimental incentives fixed, there may be a region of (θ_p, θ_m)-space where learners are too inefficient and/or too averse to extra work to provide positive labor supply. For such students, their UH traits are *bound-identified* rather than point-identified (see also Section 4.3).

3.4. **Descriptive Analyses of Website Activity.** Table 1 displays descriptive statistics of math website activity. We define *active students* as the 44.7% who completed at least two website learning tasks, *marginal students* as the 5.6% who completed one task but not a second, and *inactive students* as the remaining 49.7% who did not complete any tasks. Within the active group, the median student completed 12 learning tasks, while 4% completed all 80. Overall, we observe 749 active students who completed 16,740 learning tasks (i.e., 83,700–100,440 math problems correctly solved) across roughly 30,000 attempts and 2,000 child-hours during our 10-day sample period. Distributions of learning task completion, website time, and rate of progress illustrate striking heterogeneity: they all have medians well below the means, and standard deviations near or above the means.

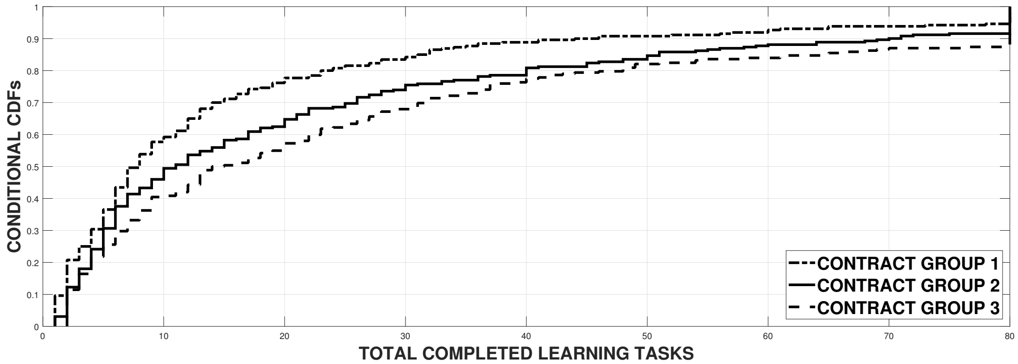
To put these numbers in perspective, first recall that website activity was extracurricular, being separate from regular schoolwork; for comparison we compiled survey data on school homework activity. Among active students, mean self-reported homework time across the pre-survey and post-survey differed by a small and statistically insignificant margin (3.97%, p-value$=0.156$).[27] These survey data also contextualize the scale of observed website activity. Assuming mathematics accounted for 25%–50% of typical homework time implies the average (median) website time among active students was an increase of 37%–74% (24%–48%) in math study activity. Active students reported 22.1% more daily homework time than marginal/inactive students (p-value 3.5×10^{-38}). We also find a statistically significant relationship between homework time and observed website activity.[28] These results suggest a meaningful link between our website data and unobservable differences between students that drive choices and outcomes.

3.4.1. *Identification Strategy.* Figure 1 and Table 1 illustrate how raw data moments pin down structural parameters. Work-time panel data identify productivity fixed effects, and mean panel length among active students is 22.35 completed tasks. Incentive randomization

[27]Our pre-survey asked about students' homework time during a "typical week;" our post-survey asked the same question, but specifically about the sample period. Table 1 reports homework time numbers averaged across pre- and post-survey reports.

[28]We estimate positive Spearman rank correlations between daily homework time and three measures of willingness to engage in extracurricular math: (binary) active status, 0.229 (p-value 1.8×10^{-21}); task accomplishment, A_i, 0.238 (p-value 4.6×10^{-23}); and time spent, T_i, 0.223 (p-value 2.5×10^{-20}). Survey responses on the frequency of missing homework deadlines at school have a negative Spearman rank correlation of -0.265 (p-value 2.6×10^{-26}) with time spent on our math website.

Figure 1. Math Website Output by Contract Group



Notes: Null hypotheses of pairwise distributional equality are rejected by the Barrett-Donald (2003) test (100,000 bootstrap samples) in favor of first-order dominance (*Group 1* vs *Group 2* p-value=0.002; *Group 1* vs *Group 3* p-value$<10^{-5}$).

serves as an instrument to tease out idiosyncratic preferences. Cost convexity implies that labor-supply CDFs should exhibit first-order dominance shifts when marginal wages rise, which is confirmed in Figure 1 since $\pi_{11} < \pi_{12} < \pi_{13}$. A formal stochastic dominance test (Barrett and Donald (2003)) reveals that the null hypotheses of pairwise equality among the three CDFs are rejected in favor of first-order dominance. This result is similar in spirit to a first-stage F-test within linear instrumental variables models. Randomization and the monotonicity property of labor supply (see Section 2.5) imply that quantile differences are interpretable as treatment effects of incentive increases for people with similar unobserved preferences. Moreover, the magnitudes of the observed quantile shifts are informative of the curvature of $c(t;\gamma_c)$. This identification logic is formalized in Sections 4.1 and 4.2 below.

4. Identification and Estimation of the Student Choice Model

The primary structural primitives of the study-leisure model are idiosyncratic productivity/motivation parameters $(\theta_{pi}, \theta_{mi})$ and the cost function $c(t)$. Additional structural parameters include $\mathcal{T}_0$, $\mathcal{T}_1$, φ, and work-time shock CDFs $F_u(u_{ia}|\theta_{pi})$. We now discuss identification and the two-stage GMM estimator used to implement our empirical strategy.

4.1. **Stage-1 Identification/Estimation: Productivity.** We follow a standard approach to estimate equation (1), using the within-child panel structure of work-time data, $\{\tau_{ia}\}_{a=1}^{A_i}$.

Assumption 4. *Study-time productivity shocks satisfy the following:*

(1) $U_{ia} \perp\!\!\!\perp \Theta_{mi}|\theta_{pi}$ *(shocks are cond. independent of motivation, given productivity);*
(2) U_{ia} *is iid across tasks* a *for each* i*; and*
(3) $\mathrm{E}[\log(U_{ia})|\theta_{pi}] = 0 \ \forall\theta_{pi}$ *(log-shocks are mean independent of productivity).*

Intuitively, the first part of the assumption means that a child's motivation trait θ_{mi} operates only on her decision to spend time on math or the outside option. Conditional on devoting time to math, she invests full cognitive resources into the incentivized task and operates at her production possibility frontier. The second part says that production times across learning tasks are unpredictable. Recall from Section 3.2.2 that learning tasks were populated with content randomly selected from our bank of math problems, so that task difficulty was impossible to predict without incurring the cost of an attempt. The third part of the assumption is similar, and implies that students of different productivity types did not differ in their ability to predict average task difficulty.[29] In Online Appendix C.1 we run a formal test proposed by Verbeek and Nijman (1992) for selectivity bias in unbalanced panel data, which Assumption 4 rules out; we fail to find evidence contradicting it. Note, however, that Assumption 4 allows for $\log(U_{ia})$ to be heteroskedastic in θ_{pi}.

Log-transforming equation (1) produces a linear-in-parameters regression,

$$\log(\tau_{ia}) = \log(\mathcal{T}_0) + \log(\mathcal{T}_1)\mathbb{1}(a=1) + \log(\theta_{pi}) - \varphi\log(a) + \log(u_{ia}),\ a=1,\ldots,A_i,\ \{i|A_i \geq 2\}, \quad (6)$$

where $\log(\theta_{pi})$ is a student fixed-effect (FE), and $\boldsymbol{B}_p \equiv [\log(\mathcal{T}_0), \log(\mathcal{T}_1), \varphi]'$ are intercept and slope terms. We estimate equation (6) by a standard panel-data differencing approach. In Online Appendix C.1 we formally define our FE estimator on the unbalanced panel of student-task time logs, and explain how established econometric theory (e.g., Wooldridge (2002), Ch. 17) implies $\sqrt{N}$-consistency and asymptotic normality of $\hat{\boldsymbol{B}}_p$, and $\sqrt{A_i}$-consistency of FE estimates $\hat{\theta}_{pi}$ (for active students) under Assumption 4.

For estimation of heteroskedastic study-time shock CDFs, $F_u(u|\theta_p)$, we first compute fitted residuals, $\hat{u}_{ia} = \tau_{ia}/(\hat{\mathcal{T}}_0\hat{\mathcal{T}}_1^{\mathbb{1}(a=1)}\hat{\theta}_{pi}a^{-\hat{\varphi}})$, $a=1,\ldots,A_i$, $\{i|A_i \geq 2\}$, and we partition the support of $\hat{\theta}_{pi}$ into 5 sub-intervals of equal length, I_j^p, $j=1,\ldots,5$.[30] Then, we split fitted residuals into 5 sub-samples $\{\{\hat{u}_{ia}\}_{a=1}^{A_i}\}_{\{i|\hat{\theta}_{pi}\in I_j^p\}}$, $j=1,\ldots,5$, and we smooth the corresponding empirical CDFs using a flexible cubic B-spline form $\hat{F}_u(u|I_j^p;\boldsymbol{\gamma}_{uj}) = \sum_{k=1}^{7}\gamma_{ujk}\mathcal{B}_{ujk}(u)$, $j=1,\ldots,5$.[31]

4.2. **Stage-2 Identification/Estimation: Labor-Supply.** Under Assumption 4, Stage-1 parameters including $\left\{\theta_{pi}, \{u_{ia}\}_{a=1}^{A_i}\right\}_{A_i\geq 2}$, $\mathcal{T}_0$, $\mathcal{T}_1$, φ, and $F_u(u|\theta_p)$ can be treated as known and fixed from the Stage-2 perspective. This provides needed tractability by reducing parameter-space dimension and computational burden. Identification of student labor-supply elasticities, including the cost function $c(t)$ and motivation types θ_{mi}, follows Guerre, Perrigne,

[29] Parts *(2)* and *(3)* of Assumption 4 are suggestively supported by the fact that it was rare for children to attempt a task, leave it permanently unfinished, and subsequently complete another task (see Online Appendix B.5 on edge-case time data).

[30] Specifically, $I_j^p \equiv \left[\min(\hat{\theta}_{pi}) + (j-1)h, \min(\hat{\theta}_{pi}) + jh\right]$, $h = (\max(\hat{\theta}_{pi}) - \min(\hat{\theta}_{pi}))/5$, $j=1,\ldots,5$. A finer partition of 10 sub-intervals of θ_p made little difference in following stages of estimation.

[31] Basis functions $\mathcal{B}_{ujk}$ are determined by the Cox-de Boor formula and a pre-specified knot vector spanning *supp*(F_u). We chose 4 knots, uniformly spaced in quantile ranks (see Appendix C.3). After constraining the endpoints this left 5 free parameters, which achieved a tight fit relative to the empirical CDFs; see Figure OS.2 (online appendix) for resulting shock densities.

and Vuong (2009) and D'Haultfoeuille and Février (2020).[32] These papers explore conditions under which discrete instruments can nonparametrically identify continuous UH (e.g., preference parameters θ_{mi}) without *a priori* functional-form restrictions on utility or the CDF of UH. They show that monotone choices and a key exclusion restriction—agents with similar UH traits making choices under different incentives (e.g., randomization)—implies that quantile differences in observed actions correspond to type-specific treatment effects.

4.2.1. *Nonparametric Identification of Labor-Supply Costs.* Suppose the cost function is indexed by a (potentially infinite-dimensional) parameter vector, $c = c(t;\boldsymbol{\gamma}_c)$. Given known incentives Π_i, productivity θ_{pi}, shock sequence $\{u_{ia}\}_{a=1}^{A_i}$, success probability F_s, and cost function parameters $\boldsymbol{\gamma}_c$, i's motivation type is uniquely determined by her choices (A_i, T_i) and the inverse solution to equation (3): $\theta_{mi} = L_T^{-1}(T_i;\theta_{pi},\Pi_i,\{u_{ia}\}_{a=1}^{A_i},\boldsymbol{\gamma}_c)$. Therefore, the only free parameters to be identified in Stage 2 are the cost function parameters $\boldsymbol{\gamma}_c$. Said differently, while the labor-supply curve $L_T(\theta_{pi},\theta_{mi},\Pi_i,\{u_{ia}\}_{a=1}^{\infty};\boldsymbol{\gamma}_c)$ is a function of arguments that differ across child index i, its shape is determined by the curvature of $c(t;\boldsymbol{\gamma}_c)$.

Let $\boldsymbol{\Omega}_{ij} = (\theta_{pi},\pi_{1j},\{u_{ia}\}_{a=1}^{A_i})$ denote i's productivity type and realized shock history through A_i, coupled with contract j. Now suppose i is assigned to contract $j>1$, and her type θ_{mi} is such that her observed labor supply $T_i = L_T(\theta_{pi};\theta_{mi},\pi_{1j},\{u_{ia}\}_{a=1}^{\infty},\boldsymbol{\gamma}_c)$ is at the r^{th} conditional quantile within group j, given $\boldsymbol{\Omega}_{ij}$.[33] Since labor supply is decreasing in θ_{mi} (conditional on shocks, productivity, and incentives), and since quantile ranks are preserved under monotone transformations, we can infer that under assignment to contract $j' < j$, i's counterfactual labor supply would be the observed r^{th} conditional quantile within contract group j', given $\boldsymbol{\Omega}_{ij'}$ (holding the first and third components fixed).

Letting $r_i = G_T(T_i|\boldsymbol{\Omega}_{ij})$ (known to the econometrician), we have quantile-r_i treatment effects, $TE(r_i,\boldsymbol{\Omega}_{ij},j') \equiv G_T^{-1}(r_i|\boldsymbol{\Omega}_{ij}) - G_T^{-1}(r_i|\boldsymbol{\Omega}_{ij'})$, $j>j'$, which hold the idiosyncratic component of costs θ_{mi} fixed. Under the model, $G_T^{-1}(r_i|\boldsymbol{\Omega}_{ij'}) = L_T(\theta_{pi},\theta_{mi}(\boldsymbol{\gamma}_c),\pi_{1j'},\{u_{ia}\}_{a=1}^{\infty};\boldsymbol{\gamma}_c)$, where $\theta_{mi}(\boldsymbol{\gamma}_c)$ is the motivation type consistent with $\boldsymbol{\Omega}_{ij}$, T_i, and $\boldsymbol{\gamma}_c$. Thus, given many $(r_i,\boldsymbol{\Omega}_{ij},j')$ triples, we can uniquely identify $\boldsymbol{\gamma}_c$ from a nonparametric regression of observed quantile treatment effects $TE(r_i,\boldsymbol{\Omega}_{ij},j')$ on their model-generated analogs,

$$\widetilde{TE}(r_i,\boldsymbol{\Omega}_{ij},j';\boldsymbol{\gamma}_c) \equiv L_T(\theta_{pi},\theta_{mi}(\boldsymbol{\gamma}_c),\pi_{1j},\{u_{ia}\}_{a=1}^{\infty};\boldsymbol{\gamma}_c) - L_T(\theta_{pi},\theta_{mi}(\boldsymbol{\gamma}_c),\pi_{1j'},\{u_{ia}\}_{a=1}^{\infty};\boldsymbol{\gamma}_c). \quad (7)$$

This logic does not assume finite-dimensional $\boldsymbol{\gamma}_c$, so we have nonparametric identification of $c(t)$ (on the support of $G_T(t)$), θ_{mi} (for $A_i \geq 2$), and the conditional CDF $F_{PM}(\theta_{pi},\theta_{mi}|A_i \geq 2)$. Note that the logic above allows for mass points at the lower bounds of the labor-supply CDFs, $G_T(0|\boldsymbol{\Omega}_{ij}) > 0$. Thus, extensive-margin selection (see Table 1) is not a threat to

[32] D'Haultfoeuille and Février (2020) uses tools derived from a more general framework by Torgovitsky (2015) and D'Haultfoeuille and Février (2015).

[33] For simplicity we ignore base wage π_{0j} here, as we focus on $j \to j' < j$ counterfactuals, which *reduce* labor supply among active students. We also hold unobserved external incentives fixed, due do individual experimental randomization.

structural identification, with the caveat that individual types $(\theta_{pi}, \theta_{mi})$ are point-identified for active students and *bound-identified* for marginal/inactive students (see Section 5). ■

4.2.2. *Method of Simulated Moments (MSM) Estimator.* The main challenge is that L_T does not admit a closed-form solution, so in our application we impose a flexible, finite-dimensional form on $c(t;\boldsymbol{\gamma}_c)$ for tractability. However, nonparametric identification results by Guerre et al. (2009) and D'Haultfoeuille and Février (2020) establish that dimension restrictions on utility are not driving identification. Specifically, we specify costs as a cubic B-Spline, $\hat{c}(t;\boldsymbol{\gamma}_c) = \sum_{k=1}^{K_c+3} \gamma_{ck}\mathcal{B}_{ck}(t)$, with knot vector $\boldsymbol{\kappa}_c = \{\kappa_{c1}, \kappa_{c2}, \ldots, \kappa_{c,K_c+1}\}$ and $K_c = 7$. B-splines are known to combine remarkable flexibility with improved numerical stability over polynomials (see Bodoh-Creed, Hickman, and Boehnke (2021)).

We recover the cost function with a MSM approach that mirrors the identification argument above. For the l^{th} guess of the cost parameters $\boldsymbol{\gamma}_{cl}$, we invert students' motivation types, given their known choices, productivity, and incentives π_{1j}. Then, we repeatedly simulate their counterfactual choices under the model with alternate incentives $\pi_{1j'}$ $j' \neq j$. We then iteratively update the guess to some $\boldsymbol{\gamma}_{c,l+1}$ where the model-generated counterfactual CDFs more closely approximate the empirical CDFs displayed in Figure 1. Full technical details on our MSM estimator of costs are relegated to Online Appendix C.2. Online Appendix C.2.2 discusses $\sqrt{N}$-consistency and asymptotic normality of $\hat{\boldsymbol{\gamma}}_c$, and $\sqrt{A_i}$-consistency of motivation estimates $\hat{\theta}_{pi}$. Online Appendix C.4 details our bootstrap procedure.

4.3. **Multi-Stage Estimation Issues.** One challenge is that FE estimates have differing variances due to the unbalanced panel: higher values of A_i lead to more precisely measured $\hat{\theta}_{pi}$. Stage-2 estimation takes Stage-1 estimates as inputs; to account for the multi-stage propagation of sampling variability, we compute Stage-1 and Stage-2 standard errors via the bootstrap, a common coping technique for this issue. In Section 5 (Stage 3), we use student FEs as an outcome variable in a Tobit regression to decompose external influences on productivity and motivation. We weight contributions to the Tobit objective function by the reciprocal of $Var(\hat{\theta}_{pi})$ for efficiency. A Tobit approach is well-tailored to incorporate both level information from point-identified UH estimates for active students and bound information for non-active students. Although the conditional joint distribution of UH, $F_{PM}(\theta_{pi}, \theta_{mi}|A_i \geq 2)$, is *nonparametrically* identified in Stage 2, we adopt a joint log-normal assumption in Stage 3 in order to recover the unconditional joint distribution of UH. Tobit estimates, combined with student covariates, allow us to compute conditional mean projections of $(\theta_{pi}, \theta_{mi})$ for non-active students, by integrating over $f_{PM}(\theta_{pi}, \theta_{mi}|\boldsymbol{X}_i, A_i < 2)$.

In Section 6 (Stage 4) we estimate skill production technology (equation 5) using student FE estimates/projections as regressors. There, we implement three measures to cope with sampling variability from Stages 1–3: (i) we compute empirical Bayes shrunk forecasts of student FEs using bootstrapped standard errors; (ii) we estimate skill technology via Feasible Generalized Least Squares to address heteroskedasticity issues that arise from estimation on an unbalanced panel in Stage 1; and (*iii*) we compute heteroskedasticity-robust standard errors and test statistics.

4.4. **Structural Estimates.** Structural parameter estimates and bootstrapped confidence intervals (400 bootstrap samples) are in Table 9 in Appendix A. These include B-Spline weights $\{\gamma_{c1}, \gamma_{c2}, \ldots, \gamma_{c10}\}$, the first two of which, γ_{c1} and γ_{c2}, are pinned down by the boundary conditions and therefore have zero sampling variance.

4.4.1. *Interim Productivity Evolution.* The experience-curve estimate, $\hat{\varphi} = 0.0788$, is positive and statistically different from zero (95%CI= $[0.074, 0.082]$). Wright's law implies $\rho_i(2a) = \rho_i(a)2^{-\varphi}$ for any labor-supply benchmark a. The expression $1 - 2^{-\varphi}$ is often referred to as the "progress ratio" (Hirsch (1956)) as it represents the fraction by which production time is reduced for the $(2a)^{\text{th}}$ unit of work, relative to production time on the a^{th} unit. Our φ estimate implies a progress ratio of 0.053, meaning that with each doubling of math activity, a child's typical production time on the $(2a)^{\text{th}}$ math assignment is 94.7% of the typical production time on her a^{th} assignment. While this effect may seem relatively small, as doubling within-period output would require large incentive shifts, it is informative of productivity evolution over time. It implies an estimated *school-year progress ratio* (SYPR) of 0.199.[34] In other words, at the end of a 36-week school year, a student's typical work time on a math assignment will be 80.1% of what it was at the beginning of the year.

4.4.2. *Heterosketastic Work-Time Shocks.* Figure OS.2 in the online appendix presents production-time shock densities. We find that productivity shock distributions for less productive (i.e., higher θ_p) students are mean-preserving spreads of the corresponding shocks for their more productive (lower θ_p) counterparts. For the middle quintile of θ_p types, the 90-10 interval of U is $[0.558, 1.893]$ with a median of roughly 1, meaning that typical task completion times range from 44% below to 89% above the median completion time.

4.4.3. *Cost Schedule.* Figure 2 plots the estimated cost function $C(t; \hat{\theta}_m, \widehat{\boldsymbol{\gamma}}_c)$, scaled to the median value of θ_{mi} among active students. Costs are precisely estimated for relatively low values of time commitment, while bootstrapped confidence bands widen for higher values where time-choice data are less plentiful. Figure 6 (Appendix A) depicts goodness of fit by comparing empirical and model-generated labor-supply CDFs by contract group. Overall, the structural model does well at matching patterns in the data, especially for contract group 2, where the richest set of counterfactual comparisons are available.

We estimate a high degree of cost curvature; Figure 2 labels cost levels at regular intervals to illustrate this point. The depicted child chose a total time commitment to our offered

[34]Recall that our sample period included 2 weeks of tracked study activity. If we think of doubling a child's learning activity across 2-week periods, then each 36-week school year entails 4.0625 doubling cycles: 2 weeks to 4 weeks, then 8, 16, 32, and 1/16th of a final doubling cycle at 36 weeks. Thus, the SYPR is $1 - 2^{-4.0625\hat{\varphi}} = 0.199$. In this calculation, there is the caveat that we can only observe math activity on our website. If we assume that unobserved baseline math work for school is roughly proportional to website activity among active students, then the 19.9% productivity gain estimate would remain unchanged.

Figure 2. Time Supply Cost & Marginal Cost Estimates

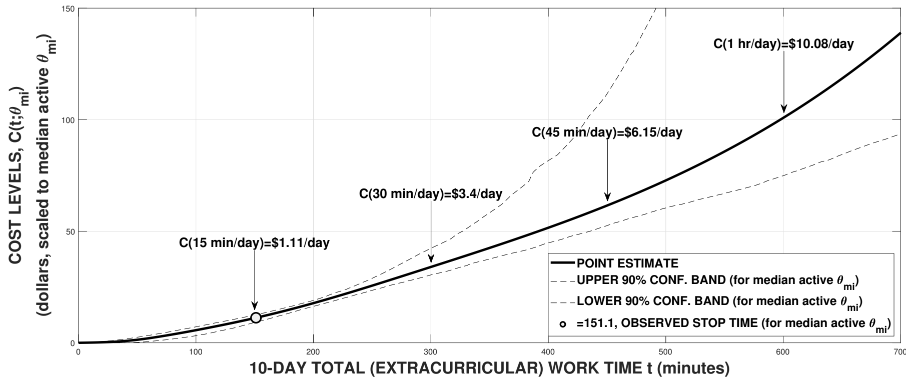
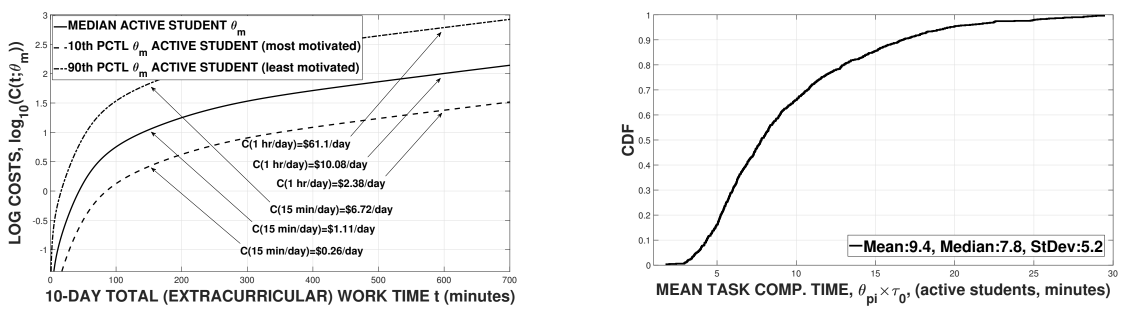


Figure 3. Motivation/Productivity Heterogeneity



extracurricular math activities of 151.1 minutes, or just over 15 extra minutes per day over our sample period. At this level of additional math activity, this child incurred a daily utility cost of just over \$1.11 per day. Doubling math-time allocation roughly triples costs, and an increase to 1 hour/day raises utility costs by an order of magnitude.

These numbers mask subtle agency issues within educational contexts. As an intervention to reach a desired competency standard, an omniscient principal could make a take-it-or-leave-it zero-surplus offer to the individual student depicted in Figure 2 of \$10.08/day to induce them to an extra hour of daily math study (or \$1.11/day for an extra 15 minutes, etc.). However, such targeted offers would require that the principal observe students' individual private information. Without it, at best the principal could offer a common incentive across observable groups of students (i.e., second-degree price discrimination, based on classrooms, program participants, etc.). From a practical policy perspective, depicted cost levels are deceptively low: they represent compensating differentials to make the child whole for giving up extra leisure, rather than the cost of solving the adverse-selection problem of motivating children to voluntarily supply time to study, accounting for individual optimal choice. For example, with the experimental piece-rate incentives based on observable output, the child in Figure 2 completed 28 learning tasks in 15.1 minutes/day and earned a surplus of \$2.80/day, or a 250% markup *above* the zero-surplus transfer of \$1.11/day.

4.4.4. *Motivation and Productivity Heterogeneity.* The left panel of Figure 3 illustrates cross-student cost variation. The figure depicts log-cost schedules scaled to θ_m types at the 10^{th} percentile, median, and 90^{th} percentile of active students. We find broad heterogeneity in willingness to supply time to learning activity: the 10-90 range (conditional on active status) entails a 25-fold increase in labor-supply costs for a fixed time commitment t. This heterogeneity only adds to the challenges of the information-constrained principal: offering sufficient uniform incentives to entice the 90^{th}-percentile θ_m type to study more will elicit large and costly responses by students who are much more motivated. Providing lower uniform incentives that are only sufficient to entice the 10^{th}-percentile types will evoke little or no labor-supply response from the rest of the population. The right panel of Figure 3 plots a productivity CDF: the 10-90 range entails a 4-fold change in median task completion time or, equivalently, a 4-fold range of hourly compensation, holding piece-rate incentives fixed. In ongoing work, Cotton et al. (2025) adopt a market-design approach to explore these agency issues and related education policy counterfactuals.

5. Building Blocks of Motivation & Productivity

This section presents Stage 3 estimation, which studies how external factors impact individuals' learning-relevant traits. Motivation heterogeneity may be driven by internal "psychic" costs of working on math, or by external factors (e.g., quality/variety of leisure activities) that shape the opportunity costs of time. Productivity differentials may reflect various internal factors (e.g., cognitive differences or baseline math skill), and external factors (e.g., instructional quality, family support, or home learning resources). The analysis considers how much variation in productivity/motivation is explained by external factors, and the potential role for education policy to influence the type of learner a child becomes. To do so, we model θ_p and θ_m as comprising internal and external components as follows:

$$\log(\theta_{pi}) = \boldsymbol{X}_{pi}\boldsymbol{\beta}_p + \eta_{pi}, \quad \text{and} \quad \log(\theta_{mi}) = \boldsymbol{X}_{mi}\boldsymbol{\beta}_m + \eta_{mi}, \tag{8}$$

where $\boldsymbol{X}_{pi}$ and $\boldsymbol{X}_{mi}$ contain student covariates and (η_{pi}, η_{mi}) represent the idiosyncratic component of student i's unobserved traits $(\log(\theta_{pi}), \log(\theta_{mi}))$. One obstacle to overcome is truncation of the outcome variable: the left-hand variables are known only for active students. While the conditional type distribution $F_{PM}(\theta_{pi}, \theta_{mi}|A_i \geq 2)$ is nonparametrically identified (see Section 4.2), it will be useful here to assume a parametric form in order to recover the *unconditional* type distribution $F_{PM}(\theta_{pi}, \theta_{mi})$.

Assumption 5. *Residual log-productivity and log-motivation are bivariate normal distributed, $(\eta_{pi}, \eta_{mi}) \sim \Phi(\mathbf{0}; \boldsymbol{\Sigma}_i)$, with variance-covariance structure depending on race and gender:*

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{pi}^2 & \sigma_{pmi} \\ \sigma_{pmi} & \sigma_{mi}^2 \end{bmatrix}, \ \sigma_{ji} = \sigma_{j0} + \sigma_{j1} fem_i + \sigma_{j2} black_i + \sigma_{j3} hisp_i, \quad j = p, m, pm.$$

Structural estimates allow us to bound marginal/inactive (θ_p, θ_m) types using a selection threshold, $\underline{\Theta}_m(\theta_p; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c)$, which is the marginal motivation type that is willing to complete $A_i \geq 2$ tasks (on average), given productivity θ_p and contract-j incentives, and $\underline{\Theta}_p(\theta_m; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c) = \underline{\Theta}_m^{-1}(\theta_p; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c)$.[35] These thresholds imply the following:

$$A_i < 2 \quad \Rightarrow \quad \boldsymbol{X}_{pi}\boldsymbol{\beta}_p + \eta_{pi} \geq \log\left(\underline{\Theta}_p(\boldsymbol{X}_{mi}\boldsymbol{\beta}_m + \eta_{mi}; \pi_0^i, \pi_1^i, \widehat{\boldsymbol{\gamma}}_c)\right). \tag{9}$$

Assumption 5 with (8) and (9) facilitate a bivariate Tobit Maximum Likelihood estimator,

$$\begin{aligned}\left[\hat{\boldsymbol{\beta}}_p, \hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\Sigma}}\right] = \operatorname{argmax}\Bigg\{ & \textstyle\sum_{i=1}^N \mathbb{1}(A_i \geq 2)\omega_{di} \log\left[\phi\left([\log(\theta_{pi}) - \boldsymbol{X}_{pi}\boldsymbol{\beta}_p, \log(\theta_{mi}) - \boldsymbol{X}_{mi}\boldsymbol{\beta}_m]; \boldsymbol{\Sigma}_i\right)\right] \\ & + \mathbb{1}(A_i < 2)\omega_{di} \log\left[\iint_{\eta_m + \boldsymbol{X}_{mi}\boldsymbol{\beta}_m > \log\left[\underline{\Theta}_m(\exp(\boldsymbol{X}_{pi}\boldsymbol{\beta}_p + \eta_p); \pi_0^i, \pi_1^i, \widehat{\boldsymbol{\gamma}}_c)\right]} \phi([\eta_p, \eta_m]; \boldsymbol{\Sigma}_i)\, d\eta_p d\eta_m\right]\Bigg\},\end{aligned} \tag{10}$$

with inverse-variance weights $\omega_{di} = \frac{1}{Var(\hat{\theta}_{pi})}$ if $A_i \geq 2$, and $\omega_{di} = \min\{\omega_{dj} | A_j \geq 2\}$ if $A_i < 2$. We compute selection probability in the second line by simulation; see Online Appendix C.5 for technical implementation details. In order to attach a causal interpretation to estimates of the parameters of equation (8), we require the following assumption on the error structure:

Assumption 6. $E[(\eta_{pi}, \eta_{mi}) | \boldsymbol{X}_{pi}, \boldsymbol{X}_{mi}] = \mathbf{0}$.

In words, this essentially assumes that our available student covariates adequately proxy for systematic external components of productivity/motivation, with residual omitted variable bias from other external factors playing a negligible role.[36] After presenting results we return to a discussion on the plausibility of this assumption in Section 5.1.5 below.

5.0.1. *Student Covariates.* $\boldsymbol{X}_{pi}$ contains an intercept and the following: indicators for *gender*, *race*, *age cohort*, and *school district*; two variables for *#adult academic helpers* and *#peer academic helpers*, defined as people who regularly assist child i with schoolwork; and two socioeconomic variables—*mean household income* and *fraction minors with no health insurance*—measured at the neighborhood (Census Block Group) level, of which there are roughly 160 across our 3 school districts. These last two variables proxy for affluence and developmental resource deprivation, respectively, but since they are not measured individually they may also proxy for neighborhood-level influences (e.g., peer/adult social networks). $\boldsymbol{X}_{pi}$ also includes three controls for access to education resources: a dummy for *no home internet*, and two variables *mobilefrac* and *tabletfrac*, being fractions of website page-loads from a smartphone or tablet device (desktop/laptop is the omitted category). As a robustness check we add four additional controls from a parent survey (see below).

Covariates for motivation $\boldsymbol{X}_{mi}$ comprise the same variables plus 13 more. The idea is that θ_{mi} represents a child's willingness to shift time away from the best outside use, a function of both direct costs and opportunity costs. The first four additional motivation

[35] Thresholds can be estimated as the northeast boundary of the convex hull of $(\hat{\theta}_{pi}, \hat{\theta}_{mi})$, given $A_i \geq 2$ and contract $j \in \{1, 2, 3\}$.

[36] In practice the error terms in equation (8) are partly composed of finite sampling error in $\widehat{\log(\theta_{pi})}$ and $\widehat{\log(\theta_{mi})}$ from previous stages of estimation. Since covariates $(\boldsymbol{X}_{pi}, \boldsymbol{X}_{mi})$ played no role in structural estimation this does not violate Assumption 6.

controls measure internal attitudes toward math—indicators for whether the child reported math as their *favorite* academic subject or *least favorite*—and two survey-elicited indices for *extrinsically motivated* and *intrinsically motivated* mindset. The remaining nine controls measure variety and quality of a child's non-math time uses: indicators for enrollment in organized *sports*, *music*, or *clubs*; *fraction leisure time in adult-supervised activity*; *#video gaming systems* at i's home; a dummy for *parental permission for gaming on weekdays*; *mean weekday recreational internet time*; *mean daily recreational screen time*; and *mean daily regular schoolwork time.* These variables are summarized in Table 7 (Appendix A). Small fractions of some survey observations are missing, between 5% and 9% depending on which variable (see Table 7). To cope with this problem, we apply standard techniques for "Regression with missing X's," as surveyed by (Little, 1992), which essentially amount to imputation of missing regressor values using projections based on available regressors.[37]

As a robustness check on our primary Tobit specification, we add a final set of covariates from a parent survey: $Involved\ Parent$, a dummy for self-reported average daily time spent with the child on schoolwork of ≥ 2 hours; $BigFamily$, a dummy for 3+ children living in the household; and two birth-order covariates, $Middle\ Child$ and $Youngest\ Child$ (the omitted category is *Oldest/Only Child*). The main shortcoming of the parent survey is that responses are only available for a 20% subsample (see Table 8, Appendix A). When incorporating these variables, we use the same imputation techniques to replace missing values for the rest of the sample. Assuming that data are missing at random (conditional on other available covariates), this approach limits the statistical power of the parent survey variables—the larger the available subsample, the more power—rather than introducing bias.

5.1. **Empirical Results: Student Type Decomposition.** Tables 2 and 3 report Tobit regressions of student types on covariates. Specification (1) includes only gender, race, age, and neighborhood-SES controls; (2) introduces school effects, and the remaining columns add increasing sets of additional controls to probe for robustness of the main effects of interest. For interpretability, the tables report *average SD Effects*, defined for continuous control variable Z as $E\left[\frac{\partial \log(\theta_{ji})}{\partial z_i}\right] \frac{SD(Z)}{SD[\log(\theta_{ji})]}$, $j = p, m$, and similarly for categorical Z, where the derivative is replaced with a difference (relative to the omitted category).

5.1.1. *Interpreting $\psi_i(a)$: Opportunity Costs vs Direct Costs.* We first consider whether $\hat{\theta}_{mi}$ estimates primarily represent fundamental motivation heterogeneity, or external factors that affect opportunity costs of time. If students' time constraints drive average willingness to study, then coefficients on outside time-use variables in Table 3—*Reg Study Time*, *Recr Screen Time*, *Extracurricular*, *Gaming/Surfing*, and *Home Connectivity*—should be positive.

[37]To fix ideas, suppose $\boldsymbol{X} = \{1, X_{1i}, X_{2i}, X_{3i}\}_{i=1}^{N}$, but the value of X_{1i} is missing for some i. For imputation, we first form data matrices $\tilde{\boldsymbol{X}} = \{1, X_{2j}, X_{3j}\}_{j=1}^{J}$ and $\tilde{\boldsymbol{Y}} = \{X_{1j}\}_{j=1}^{J}$, where J is the maximal number of observations for which the values of all three regressors are available. Second, we project $\tilde{\boldsymbol{Y}}$ onto $\tilde{\boldsymbol{X}}$ to get $\hat{X}_{1i} = [1, X_{2i}, X_{3i}]\hat{\boldsymbol{B}}$ for each i with X_{1i} missing. We also include some interactions and additional variables not in Tables 2 and 3 to aid imputation (see supplemental materials).

A positive coefficient on *Reg Study Time* would be consistent with low measured motivation being driven by above-average regular coursework duties.[38] However, the coefficient on *Reg Study Time* is small, negative, and statistically insignificant. Moreover, a Wald test fails to reject the joint exclusion of all 12 outside time-use controls (p-value≈0.93).

For interpretation of this result, it is useful to consider a model of holistic time allocation choices (e.g., extracurriculars, non-math schoolwork, leisure, etc.). For opportunity cost of outside options to drive website labor-supply, the average child would need to have a non-negligible Lagrange multiplier on her time constraint. To see why, consider an exhaustive menu of $K+1$ activities, where T_{0i} is time on our math website, and $\boldsymbol{T}_i = (T_{1i}, \ldots, T_{Ki})$ is a vector of outside time choices. Suppose that utility is quasi-linear in monetary surplus, $U(T_{0i}, \boldsymbol{T}_i) = u(\boldsymbol{T}_i\boldsymbol{\beta}_{ui}) + M(T_{0i}; \theta_{pi}, \Pi_i) - \theta_{mi}c(T_{0i})$, where $M(T_{0i}; \theta_{pi}, \Pi_i)$ is expected earnings as a function of T_{0i}. Finally, child i maximizes utility subject to a time constraint: $\max_{\{T_{0i}, \boldsymbol{T}_i\} \in \mathbb{R}_+^{K+1}} U(T_{0i}, \ldots, T_{Ki})$ *s.t.* $T_{0i} + \cdots + T_{Ki} = \overline{T}$. The FOC for math-website time is $-\theta_{mi}c'(T^*_{0i}) + M'(T^*_{0i}; \theta_{pi}, \Pi_i) = \lambda^*_i$, where M' is the left-hand side of (4), and outside choices $\boldsymbol{T}^*_i$ influence T^*_{0i} through λ^*_i, the Lagrange multiplier on i's time constraint. Failure to reject a joint exclusion of the 12 outside time-use controls suggests that the average 5^{th}-/6^{th}-grader has a relatively low shadow value of time, λ^*_i, and is not highly constrained after other daily activities are accounted for. Instead, results from Table 3—using data on active, marginal, and inactive students—suggest that the average child bases study decisions primarily on direct psychic costs (e.g., work aversion, burnout, etc.) rather than on opportunity costs.

5.1.2. *Internal Influences on Motivation.* Psychic costs of effort have long been theorized in the education literature (e.g., Kurzban, Duckworth, Kable, and Meyers (2013)). We find suggestive evidence that preferences for math as a favorite subject raise motivation by 0.16 SD, while listing math as one's least-favorite subject reduces motivation by 0.16 SD. These estimates are somewhat noisy (joint p-value=0.208), but the large magnitudes hint at the origins of psychic study costs. In Table 3, being either more extrinsically minded or more intrinsically minded are both strong indicators of responsiveness to external academic incentives.[39] This contributes to recent empirical evidence of a synergistic role for intrinsic and extrinsic incentives (e.g., Kremer, Miguel, & Thornton, 2009; Hedblom et al., 2022), rather than a conflicting role as previously thought (e.g., Gneezy & Rustichini, 2000; Bénabou & Tirole, 2003; Leuven, Oosterbeek, & van der Klaauw, 2010). This also quantifies another aspect of psychic costs: a one SD increase in intrinsic mindset score raises motivation by more than enough to overcome aversion to math as one's least favorite subject.

[38]If so, the researcher could remove spurious apparent motivation by computing residual $\hat{\theta}_{mi}$, net of study time.

[39]For intrinsic/Extrinsic mindset scores, two questions each on the pre- and post-survey asked about one's most salient motivation for completing school-related work. Two external motivation choices were listed with two intrinsic choices, and a fifth "none of the above" option. We counted the number of corresponding responses (up to four per category) and standardized the resulting scores. The fifth option means a student may be coded as extrinsically minded, intrinsically minded, both, or neither.

5.1.3. *Year-on-Year Evolution of Latent Child Traits.* Although our model is one of short-term choices, pooling 5^{th}- and 6^{th}-graders in the study allows us to measure year-on-year evolution of (θ_p, θ_m) within the sample population. Table 3 (*Grade-5* coefficient) indicates that 5^{th}- and 6^{th}-graders are on average indistinguishable in their motivation for math activity. In Table 2 we find that 5^{th}-graders are less productive, on average, by 33.2% of a standard deviation, after controlling for the full set of covariates. This result is remarkably stable across all specifications. A standard deviation of task-completion time was 7.38 minutes and the mean (median) was 10.33 (7.84), which implies a *calendar-year progress ratio* (CYPR) of 0.237 (0.313). The SYPR calculated in Section 4.4.1 (0.199, using $\hat{\varphi}$) accounts for roughly 84% (64%) of this number. There are natural reasons to expect SYPR$\leq$CYPR, aside from the latter measuring 12 months of child development while the former accounts for only 9: the remaining 16% (36%) of the average (median) CYPR could be due to spillovers from other developmental pursuits, as suggested by Cunha et al. (2010).

5.1.4. *School District Effects.* We wish to address two questions: after controlling for other contextual factors, (i) do school quality differentials make some kids more motivated learners, and (ii) do they make some kids more productive learners? For the first question, the answer is no: estimated school effects on $\log(\theta_m)$ in Table 3 are small and statistically insignificant. Regarding the second question, in Table 2 one's school enrollment predicts significant shifts in time required to complete learning tasks, conditional on other covariates. Descriptive evidence in Table OS.4 hints that District 1's inputs—higher funding per student, larger fraction of budget devoted to instruction, more qualified/better paid faculty/administrators—are more advantageous than District 2's and District 3's. School value-added estimates in Table 2 confirm this expectation: switching from District 1 to District 2 or District 3 predicts a productivity reduction of 0.19 SD and 0.62 SD, respectively. School productivity effects are jointly significant in our main specification (4) (and robust to inclusion of parent survey controls in (5)), with individual school effects being statistically significant at the 99% level. Coefficient magnitudes are fairly stable across four separate specifications, suggesting robustness to inclusion of a rich set of childhood contextual factors.

Our Tobit results speak to a classic question of whether better outcomes at higher-performing schools are due to treatment by more advantageous school inputs, or due to selection of students onto their rolls with better resources and preparation. We find evidence for both explanations: while higher-performing schools benefit from significant advantageous selection on observables $(\boldsymbol{X}_{pi}\boldsymbol{\beta}_p, \boldsymbol{X}_{mi}\boldsymbol{\beta}_m)$ and on unobservables (θ_p, θ_m), they also appear to exert their own advantageous influence on productivity as well.

5.1.5. *Threats to Causal Interpretation of School Effect Estimates.* Our identification strategy for school value-added (VA), based on field-experimental data, differs from existing approaches. A typical study on school VA would use observational data with a large sample of schools, and outcomes (e.g., exam scores) often aggregated to the classroom or school level (e.g., see (Ahn et al., 2022), (Luccioni, 2024)). Some studies use a similar many-schools approach in combination with some source of plausibly exogenous variation to tease apart school VA from selection on UH; e.g., Dale and Krueger (2002) and Mountjoy and Hickman (2020). Other work by Abdulkadiroglu et al. (2020) combines standard VA methods with equilibrium theory to tease apart causal VA from preferences. In our case, we have a small set of school districts but novel student-level observables—real-time tracking of home-study activity and behavioral responses to experimental incentives—that facilitate identification of UH in preferences and study productivity, independently of school assignment. This, in turn, facilitates analysis of several forms of school VA in Sections 5 and 6.

To illustrate our VA identification strategy, we contrast it with standard VA methods, surveyed by Koedel et al. (2015). A prototypical approach to VA posits a model like

$$Y_{ist} = \beta_0 + Y_{ist-1}\beta_1 + \boldsymbol{X}_{ist}\boldsymbol{\beta}_2 + \epsilon_{ist}, \quad \epsilon_{ist} = \xi_i + \zeta_s + e_{ist}, \tag{11}$$

where Y_{ist} is an academic outcome (e.g., test score) for student i in school s and period t, $\boldsymbol{X}_{ist}$ is a vector of student characteristics, and the unobserved error is expandable to include ξ_i and ζ_s, which represent the impacts of student ability and school VA on the outcome Y_{ist}. The canonical problem for causal VA identification is one of selection: equilibrium residential sorting patterns imply that students with more advantageous traits congregate at schools with more advantageous traits, creating correlation between unobserved ξ_i and school assignment. Importantly though, the economic interpretations of residuals in (8) and (11) differ: ξ_i is the effect of latent student traits on academic outcome Y_{ist}, while residuals (η_{pi}, η_{mi}) are unexplained components of latent student traits themselves.

Thus, the primary considerations in attaching a causal interpretation to school effects in Tables 2 and 3 are two-fold: first, controlling for factors that drive residential sorting patterns and also affect productivity/motivation; and second, controlling for other relevant sources of omitted variable bias. For the first concern, regressors that control most directly for school-related residential sorting are our two neighborhood-SES controls, though other potentially relevant controls include *#adult helpers*, *involved parent*, the 3 extracurricular enrollment dummies, and *fraction supervised leisure time*. These last 6 variables proxy for the possibility that more invested parents may cluster at higher VA schools. Rothstein (2006) and Abdulkadiroglu et al. (2020) find evidence that parents' residential sorting and school choices appear to be based on factors like local peer quality rather than school VA. These results are consistent with the idea that neighborhood-SES controls in specifications (1)–(5) adequately proxy for parents' residential sorting choices, and could explain why school productivity effects are robust to inclusion of additional covariates.

As for the second concern—general residual omitted-variable bias—as in other studies there is no way to completely rule this out. The best we can hope for is to control or proxy for what we believe are the most important sources of potential omitted variable bias. This involved designing our data collection measures to gather as much relevant information as possible, given constraints involved in surveying adolescents during their school day. In order for our estimated school VA effects to be largely or mostly attributable to omitted variable bias, there would have to exist some aspect of student productivity that is both strongly correlated with school assignment, and also not well proxied for by the various demographic, neighborhood, and household controls in Tables 2 and 3.

5.2. **Demographics.**

5.2.1. *Racial Differences.* Many well-meaning researchers and practitioners believe that minority students from less affluent backgrounds (and worse academic outcomes), primarily struggle with a lack of motivation or engagement with academics. If anything, our Table 3 results suggest the opposite. Black and Hispanic students in our sample may be *more* conditionally motivated, though these effects are noisy (race joint p-value=0.32 in (4)). Black/Hispanic students being academically as or more motivated is also supported by patterns in our raw data: they are more likely to report math as their favorite subject than White/Asian students (43.5% vs 30.2%, zero-difference p-value=3.9×10^{-8}), and also exhibit higher levels of intrinsic motivation, on average (zero-difference p-value=0.005).

In our data, the main difference that could explain achievement gaps is a large racial productivity differential. It arises primarily because: (i) Black/Hispanic student enrollment is concentrated in different school districts, and (ii) there are residual factors proxied for by race, which persist after controlling for gender, school, family, and neighborhood factors. The conditional race gap in productivity is substantial, with $\log(\theta_p)$ being nearly a full SD higher for Black students and 2/3 SD higher for Hispanic students. This productivity gap dominates the estimated motivation advantage for Black/Hispanic students, and explains why they completed fewer website tasks (Figure 5, Appendix A).

5.2.2. *Gender Differences.* We find evidence of a female advantage in motivation and a disadvantage in study-time productivity. In Table 2, the rate of progress through learning tasks is 21% (main specification (4)) of a SD lower for females. Though smaller than the race gap, this difference is non-negligible. On the other hand, results in Table 3, suggest that female adolescents are also more motivated, with willingness to spend time on math being roughly 40% SD higher, on average. For academic labor-supply, the latter effect dominates, explaining why females complete more website learning tasks (Figure 5, Appendix A).

5.3. **Variances & Covariance of Latent Traits.** The Tobit estimator allows us to gauge the correlation between productivity and motivation for the population at large, including both active students ($A_i \geq 2$), and inactive students ($A_i < 2$). Table 4 presents estimated standard deviations and correlations of log-learning types, $(\log(\theta_p), \log(\theta_m))$, for Tobit specifications (4) and (5). We also present correlations broken down by the predictable component, $(\boldsymbol{X}_{pi}\boldsymbol{\beta}_p, \boldsymbol{X}_{mi}\boldsymbol{\beta}_m)$, and the idiosyncratic component, (η_{pi}, η_{mi}). The striking result is that type correlations (averaged across races/genders) are small and statistically indistinguishable from zero. Accounting for sampling variability, we can rule out moderate or large positive correlations between adolescent learners' productivity and motivation; in fact, point estimates lean toward a small *negative* correlation.

To put this result in context, our sample, which contains many high-achieving students, includes relatively *few* individuals who are both highly productive (i.e., low θ_p) and highly motivated (i.e., low θ_m) for math study. It is more common for high achievers to be *either* highly productive *or* highly motivated, but not both. This result challenges common labels we use for high-achievers as "gifted and talented": indeed, many such students are not particularly gifted with an ability to quicly progress through learning tasks, but instead are exceptionally hard working. This novel result provides reason for optimism among educators and policy makers struggling to provide education services to a heterogeneous population: a child need not have it all to attain a high degree of academic success. On the other hand, it also highlights that one-size-fits-all approaches may be ineffective when developing solutions for students who fail to meet competency standards.

5.4. **UH Projections for All Students.** In Section 6 we will control for UH in motivation and productivity when estimating skill production technology, but to do so there are two issues to overcome. First, since we only have point estimates of $(\theta_{pi}, \theta_{mi})$ for active students $(A_i \geq 2)$, there is a missing variables problem. This is straightforward to address using Tobit estimates: for marginal/inactive students $(A_i<2)$ we compute conditional expectations,[40]

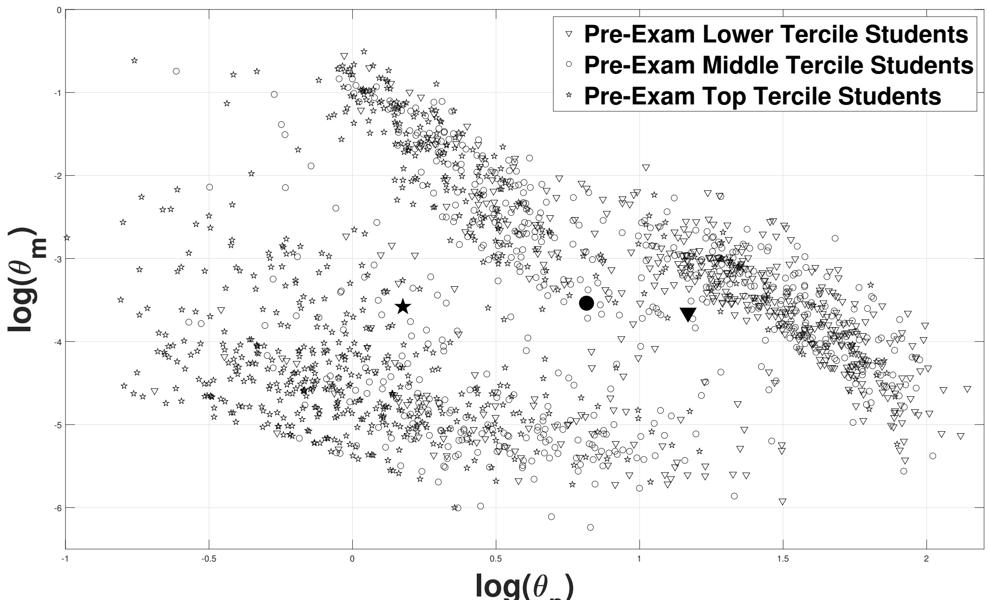
$$\left(\widehat{\log(\theta)}_{pi}, \widehat{\log(\theta)}_{mi}\right) = E\left[\left(\log(\theta_p), \log(\theta_m)\right)\middle| \boldsymbol{X}_{pi},\ \boldsymbol{X}_{mi},\ A_i<2,\ (\pi_0^i, \pi_1^i);\ \hat{\boldsymbol{\beta}}_p, \hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\Sigma}}_i\right].$$

The second issue is an errors-in-variables problem from sampling variability in student fixed-effect estimates. We compute Empirical Bayes (EB) forecasts of (θ_p, θ_m) to reduce attenuation bias by shrinking fixed effects toward the mean in proportion to the individual noise in each estimated fixed effect. This approach has a long history in the literatures on school quality (e.g. Kane & Staiger, 2002) and teacher value-added (e.g. Jacob & Lefgren, 2008). One standard procedure (e.g. Morris, 1983; Abdulkadiroglu et al., 2020) is to assume a normal prior over the true fixed effect, $\log(\theta_{ji})$, and the estimation residual, e_{ji} for $j=p,m$. This implies a shrinkage factor of $\lambda_{\theta ji} = \nu^2_{\theta j} / \left(\nu^2_{\theta j} + \nu^2_{\theta ji}\right)$, where $\nu^2_{\theta j}$ is the estimated variance of true $\log(\theta_{ji})$, and $\nu^2_{\theta ji}$ is the estimated sampling residual variance on $\widehat{\log(\theta_{ji})}$ for individual i's trait $j=p,m$. This results in the following EB estimates for student traits to be used below: $\widehat{\log(\theta_{ji})}_{EB} = \lambda_{\theta ji}\widehat{\log(\theta_{ji})} + (1-\lambda_{\theta ji})\frac{\sum_{i=1}^{N}\widehat{\log(\theta_{ji})}}{N}$, $j=p,m$. Figure 7 (Appendix A) plots a comparison of raw type estimates versus their EB shrunk forecasts.

5.5. **Do Low-Performing Students Lack Motivation?** Figure 4 presents a scatter-plot of empirical Bayes estimates of $(\log(\theta_{pi}), \log(\theta_{mi}))$ for all students in the sample. It separates them by pre-test terciles—triangles for the lower tercile, circles for the middle, and stars for the upper—and we include large shaded shapes to mark the average motivation and productivity values within each tercile. We see wide variation in productivity-motivation pairs across students, but a comparison of the means reveals little difference in average motivation across the three exam-score terciles. In other words, *lower performing students*

[40]This in the spirit of standard methods for regression with missing regressors (see survey by Little (1992)).

Figure 4. Student Productivity and Motivation by Pre-Test Tercile



Notes: For active students ($A_i \geq 2$), plotted points are EB forecasts of student traits based on observed choices. For marginal/inactive students ($A_i < 2$), plotted points are conditional means of $(\log(\theta_p), \log(\theta_m))$, given covariates and Tobit parameters (Tables 2 and 3 specification (4)). Large bolded shapes are within-group means.

are not less motivated, on average, than higher performing students. Rather, the strongest distinguishing trait across pre-test terciles is study productivity θ_{pi}.

The results in Section 5 suggest new insights on education interventions that aim to decrease gender or racial academic gaps by motivating students through incentives or information about the returns to education (e.g., Fryer (2011); Levitt et al. (2016); Gneezy et al. (2019); Cotton, Nordstrom, Nanowski, and Richert (2024)). Motivation is not the primary delimiter in various comparisons that are often discussed in debates on math education policy, including Black/Hispanic students vs White/Asian students, males vs females, and high achievers vs low achievers. Thus, policies that seek to incentivize further effort by students from marginal groups may struggle to overcome the more salient productivity disadvantages they face. Within the academic labor-supply model, more productive students complete homework in less time, before utility costs rise too high to rationally continue study. This, and cost curvature, which tends to limit labor-supply of time, may explain why conditional cash transfers to students or families for improved academic performance have often resulted in limited returns, as in Fryer (2011) and Levitt et al. (2016).

6. Estimating Skill Production Technology

Stage 4 estimation considers skill production technology (equation (5)).

6.1. **Skill Assessment Data.** Recall that we observe scores from classroom pre- and post-tests (see also Online Appendix B.3). Let κ_i^k denote i's math proficiency measured by test score S_i^k, $k \in \{pre, post\}$. Skill gains are defined as $\Delta\kappa_i \equiv \kappa_i^{post} - \kappa_i^{pre}$, and measured by changes in post-test scores, $\Delta S_i = S_i^{post} - S_i^{pre}$, over the 3-week study period. Table 5 reports descriptive statistics on S_i^{pre}, S_i^{post}, and measured skill changes ΔS_i by sub-group. Since these variables will be used to study evolution of underlying skill, we have a measurement-error problem: S_i^k is a noisy proxy for skill stock, κ_i^k, given the finite exam length (36 questions). We address this issue with an EB approach similarly as in Section 5.4. We assume test $k \in \{pre, post\}$ is a noisy measure of skill, $S_i^k = \kappa_i^k + \varepsilon_{\kappa i}^k$, where $S_i^k = \frac{1}{J}\sum_{j=1}^{J} M_{ij}$ is the sample mean of $J = 36$ *iid* draws from a Bernoulli random variable M_{ij} (the outcome of exam problem j) with success probability κ_i^k.[41] Thus, $\varepsilon_{\kappa i}^k$ is finite-sampling error of a Bernoulli sample mean, and we have $Var(\varepsilon_{\kappa i}^k|\kappa_i^k) = \frac{\kappa_i^k(1-\kappa_i^k)}{36} \equiv \nu_{ki}^2$. Shrinkage factors are thus $\lambda_{ki} = \nu_k^2 / \left(\nu_k^2 + \nu_{ki}^2\right)$, where ν_k^2 is the mean of ν_{ki}^2 across i, and EB projections are $S_{EBi}^k \equiv \lambda_{ki}\sum_{i=1}^{N}\frac{S_i^k}{N} + (1-\lambda_{ki})S_i^k$.[42] In Table 5, EB estimates of skill standard deviations $SD(\kappa_i^k)$ are roughly 37% and 38% smaller than test-score standard deviations $SD(S_i^k)$ for the pre- and post-test, respectively.

Our data highlight a substantial race gap in exam scores, consistent with evidence from other studies (e.g. Clotfelter, Ladd, & Vigdor, 2009; Hanushek & Rivkin, 2006, 2009; Bodoh-Creed & Hickman, 2024). On the pre-test, White/Asian students correctly solved 8.67 additional math problems (1.08 SD), on average, relative to Black/Hispanic students. The gender gap is smaller, at 1.39 additional math problems (0.17 SD) correctly solved for males, relative to females. The table shows preliminary evidence that extracurricular math activity on our website contributed to skill gains. Active students saw mean increases of 2.60 math problems solved on the post-test, while marginal/inactive students improved scores by only 0.65 points, on average. A two-sample t-test rejects equality of these two changes (p-value= 2.5×10^{-16}).

6.2. **Skill Technology Specification.** We assume the following form for equation (5):

$$\begin{aligned}
&\Delta\kappa_i = \kappa^{post}(\kappa_i^{pre}, A_i, T_i, \theta_{pi}, \theta_{mi}, \boldsymbol{X}_i, \varepsilon_i) - \kappa_i^{pre},\\
&\kappa_i^{post} = \Delta_{0i} + \Delta_{1i}T_i + \Delta_{2i}T_i^2 + \Delta_{3i}A_i + \Delta_{4i}A_i^2 + \Delta_{5i}(T_i \times A_i) + \varepsilon_i, \quad \text{and} \\
&\quad \Delta_{ji} \equiv \boldsymbol{V}_i\boldsymbol{\delta}_j,\ \boldsymbol{V}_i = [1, \kappa_i^{pre}, \log(\theta_{pi}), \log(\theta_{mi}), \boldsymbol{X}_i], \quad \boldsymbol{\delta}_j \in \mathbb{R}^{K+4}, \quad j = 0,\ldots,5, \quad i = 1,\ldots,N.
\end{aligned} \tag{12}$$

The most closely related paper to ours is Todd and Wolpin (2018), and equation (12) has a similar general structure to their production technology: post-investment skill stock κ_i^{post} can be de-composed as the sum of prior skill times a pass-through parameter, $(1-\delta_{01})\kappa_i^{pre}$, plus a skill growth term, $[T_i, T_i^2, A_i, A_i^2, T_i \times A_i] \times [\Delta_{1i}, \ldots, \Delta_{5i}]^\top$, which is also influenced by κ_i^{pre}. The intercept Δ_{0i} and slope parameters $\{\Delta_{1i}, \ldots, \Delta_{5i}\}$ depend on various student-specific

[41] Williams (2019a) refers to independence of M_{ij} outcomes across j as *local independence*. We ignore j-specific variation—exams were designed with randomly sequenced difficulty—and we take κ_i^{pre} to be i's unconditional pre-test success probability.

[42] For ν_{ki}^2 estimates, we use $\hat{\kappa}_i^k = S_i^k$ if $0 < S_i^k < 36$; $\hat{\kappa}_i^k = \frac{1}{2\times36}$ if $S_i^k = 36$, and $\hat{\kappa}_i^k = 1 - \frac{1}{2\times36}$ if $S_i^k = 0$.

factors, including covariates $\boldsymbol{X}_i$.[43] By including UH types $\log(\theta_{pi})$ and $\log(\theta_{mi})$ in the vector of control variables, $\boldsymbol{V}_i$, we allow them to play a dual role in skill acquisition: aside from determining labor supply (L_T, L_A), they may also directly impact the rate at which a fixed volume of work (A_i, T_i) is converted into durable skill gains. Initial skill κ_i^{pre} proxies for investment in prior periods (through Δ_{0i}), and enables a test for increasing or decreasing returns-to-scale (through Δ_{ji}, $j \geq 1$), where skill gains of a given size, $\Delta\kappa$, become more or less labor-intensive to produce from a higher skill baseline, κ^{pre}. To causally interpret skill technology parameters, we require the following:

Assumption 7. $E\big[(\kappa_i^{pre}, \boldsymbol{X}_i)^\top \varepsilon_i | \theta_{pi}, \theta_{mi}\big] = \boldsymbol{0}$ and $E\big[(A_i, A_i^2, T_i, T_i^2, A_i \times T_i)^\top \varepsilon_i | \theta_{pi}, \theta_{mi}, \kappa_i^{pre}, \boldsymbol{X}_i\big] = \boldsymbol{0}$.

Assumption 7 is weaker than analogous conditions previously imposed on error structure in the literatures on skill development (see footnote 16), IRT (e.g., Williams (2019b), Assumption 2.1), and school VA (see Koedel et al. (2015)); we only require skill growth shocks to be exogenous, *conditional on UH* in productivity and motivation.

Our structural approach with field experimental data has five key advantages relative to these prior literatures. First, it simplifies a problem of item-response data within psychometrics and econometrics (see survey by Cunha et al. (2021)): test scores have no inherent cardinal scale, so using them to proxy for skill may impose arbitrary and non-innocuous scales on the distribution of latent child traits. Here, cardinal scales of UH characteristics are predetermined and anchored to interpretable variables that determine optimal labor supply; θ_{pi} is measured by time-to-completion for learning tasks, and θ_{mi} is measured by compensating differentials for time spent on math. Second, our approach avoids measurement-error problems from using time and resource surveys to proxy for investment activity (e.g., Cunha et al. (2010); Todd and Wolpin (2018); Del Boca et al. (2024)). We control for interim investment using direct, real-time observations of web-based learning activity. Third, relative to the prototypical school VA equation (11), we can bring UH ξ_i out of the error term and explicitly control for selection on unobservables into different schools. Fourth, we need not assume scalar UH, and can model latent traits as a two-dimensional object $\xi_i = (\log(\theta_{pi}), \log(\theta_{mi}))$.

Fifth, as a result of the points above, we are able to avoid restrictive functional forms—e.g., Cobb-Douglas, CES, or trans-log technology—and we need not impose monotonicity constraints on the inputs of equation (5).[44] This innovation is not merely technical, but produces four additional benefits with simple economic interpretations. We can test whether

[43] A more familiar way to understand the skill technology equation would be to rearrange its terms as follows:

$$\kappa_i^{post} = \boldsymbol{V}_i\boldsymbol{\delta}_0 + \delta_{10}T_i + \delta_{20}T_i^2 + \delta_{30}A_i + \delta_{40}A_i^2 + \delta_{50}T_i \times A_i + \widetilde{\boldsymbol{V}}_i\widetilde{\boldsymbol{\delta}}_1 T_i + \widetilde{\boldsymbol{V}}_i\widetilde{\boldsymbol{\delta}}_2 T_i^2 + \widetilde{\boldsymbol{V}}_i\widetilde{\boldsymbol{\delta}}_3 A_i + \widetilde{\boldsymbol{V}}_i\widetilde{\boldsymbol{\delta}}_4 A_i^2 + \widetilde{\boldsymbol{V}}_i\widetilde{\boldsymbol{\delta}}_5 (T_i \times A_i) + \varepsilon_i,$$

where "tilde" versions of $\boldsymbol{V}_i$ and $\boldsymbol{\delta}_j$ omit the constant term. In other words, we regress post-test scores on covariates $\boldsymbol{V}_i$, polynomial terms $\{T_i, T_i^2, A_i, A_i^2, T_i \times A_i\}$, and a complete set of covariate-polynomial interactions.

[44] Although we specify equation (12) as a quadratic for tractability, given sufficient data one could specify it as an L-dimensional complete polynomial in (A_i, T_i), which can approximate continuously differentiable real-valued functions to any desired precision (for large enough L), as a consequence of Taylor's approximation theorem (see Judd (1998, Chapter 13)).

study time or intermediate inputs produced by study time (or both) are responsible for skill growth. We can separately measure the direct role of prior skill stock (through Δ_{0i}) and its indirect role (through $\{\Delta_{1i}, \ldots, \Delta_{5i}\}$) in producing skill gains. Our skill technology can capture latent trait interactions, including between different components of UH and between UH and external factors. And finally, it can also capture direct impacts of school VA (through Δ_{0i}), while allowing for heterogeneous VA treatment effects (through $\{\Delta_{1i}, \ldots, \Delta_{5i}\}$).

6.2.1. *Estimator and Implementation Details.* There are two remaining challenges to address. First, the feasible version of (12) has a measurement-error problem in the exam score control S_i^{pre} and the outcome S_i^{post}. Instead, we use EB projections of the outcome S_{EBi}^{post} for efficiency, and the skill control S_{EBi}^{pre} to inhibit attenuation bias. The second challenge is that unbalanced panel data from Stage 1 may induce heteroskedastic errors in equation (12); in our data, formal tests reject the null hypothesis of homoskedastic errors in all specifications. Therefore, we estimate the production parameters $\{\boldsymbol{\delta}_0, \ldots, \boldsymbol{\delta}_5\}$ via feasible generalized least squares in the familiar way (Wooldridge (2016, Chapter 8)), and we compute heteroskedasticity-robust standard errors and test statistics. Thus, standard econometric theory establishes consistency and asymptotic normality under Assumption 7.

The full covariate vector $\boldsymbol{X}_i$ contains the following variables: indicators for *gender*, *race*, *grade level*, and *school district*; two *neighborhood-SES* variables; *total #academic helpers* in a child's social network; and home resource proxies (*no home internet*, *mobilefrac*, *tabletfrac*). Four different regression specifications are summarized in Table 6. Specification **(1)** includes polynomial terms, prior skill, and school dummies. Specification **(2)** adds UH in order to gauge selection bias in school VA effects in **(1)**. Specification **(3)** adds age, race, and gender controls, and our main specification **(4)** includes the full set of controls.[45]

6.3. **Skill Production Technology Results.** For interpretability, Table 6 reports *Average Standard Deviation Effects*, defined for continuous control z as

$$Avg\ SDE_z \equiv E\left[\frac{\partial \Delta\kappa_i}{\partial z_i}\right]\frac{SD(Z)}{SD(\Delta\kappa)},$$

where the partial derivative is taken as a *ceteris paribus* change in z alone. For categorical z's, Average SDEs replace the derivative inside the expectation with a difference (relative to the omitted category). This definition reflects our focus on *skill growth* $\Delta\kappa_i$, rather than on final skill stock κ_i^{post}. It asks the question of how much extra skill growth is projected when an input rises by a standard deviation, and then normalizes the projected change by a standard deviation of observed skill growth over the sample period. By default, the Average SDE

[45] We computed estimates, standard errors, and test statistics in MATLAB using numerically stable matrix inversion methods (e.g., LU, QR, Cholesky decomposition, reliable for condition number $\leq 10^{10}$). We also estimated a fifth specification as a robustness check, which added parent survey controls. Average SD effect estimates were similar to specification **(4)**, but we do not report specification **(5)** because it produced severe numerical instability for calculation of Wald tests. This was due to the large number of controls (120), as well as parent survey variables being imputed for 80% of the sample and interacted with the five polynomial terms in equation (12). Specification **(5)** MATLAB code is included with supplemental materials.

summarizes the combined impacts of control variable z, both directly through the intercept term Δ_{0i} and indirectly through the slope terms $\{\Delta_{1i}, \ldots, \Delta_{5i}\}$ in equation (12). For certain key variables, we break out these two channels of influence in Table 5.

We use the EB estimate $SD(\Delta\kappa) = 2.991$ as an outcome benchmark, representing 3 additional math problems solved out of 36, within a timed (35-minute) exam. In regression analysis, standard deviations are often used as units of "typical" shift for random variables, but they lose that intuitive meaning in highly skewed distributions. Such is the case for labor-supply inputs, (A, T) (see Table 1 and Figure 1): SDs of $(24.3, 145.8)$ exceed the 80^{th} percentiles, and would constitute an extreme hypothetical shift for many students. Instead, we adopt the *pseudo-standard deviation*, $pSD(j) \equiv G_j^{-1}(0.5|active) - G_j^{-1}(0.159|active)$, $j = T, A$ (equivalent to the usual SD for normal data), for computing average SDEs of labor supply. A pSD for A (for T) is 8.3 learning tasks (64.1 minutes); to convert reported average pSD effects in A (T) to average SD effects, simply multiply by 2.93 (2.27).

6.3.1. *Learning Results: Prior Skill Stock.* For context, Table 6 reports mean and standard deviation of raw skill growth $\Delta\kappa_i$, and *Net Baseline learning*, defined as skill growth independent of website math study, $\hat{\Delta}_{0i} - 36 \times \kappa_i^{pre}$. In the first section of the table, the average SD effect of κ_i^{pre} is negative: raising prior skill stock by one SD reduces skill growth by 0.42 SD, or roughly 1.25 test-score points. To see why, first note from equation (12) that our object of study is $\Delta\kappa_i = \kappa_i^{post} - \kappa_i^{pre}$, which implies

$$\frac{\partial \Delta\kappa_i}{\partial \kappa_i^{pre}} = \left(\frac{\partial \kappa_i^{post}}{\partial \kappa_i^{pre}} - 1 \right).$$

Second, the positive effect of κ_i^{pre} through the intercept term implies that raising pre-test score by one SD (5.001 points) *increases* post-test score, but by a smaller amount (4.002 points). The implied 80% pass-through rate of prior skill is somewhat lower than Todd and Wolpin (2018)'s estimate of 93.7%, but within their 95% confidence interval. Moreover, we find a decreasing returns-to-scale skill technology: the negative slopes-only effect of κ_i^{pre} means that improving test scores by a fixed margin becomes more labor-intensive as one's baseline skill rises.

6.3.2. *Learning Results: Labor-Supply.* For academic labor-supply variables, recall that average SD effects are based on partial derivatives, which are *negative* for T and positive for A. This novel result implies that volume of problems solved, A_i, is the primary driver of skill growth, $\Delta\kappa_i$, while total time spent, T_i, tempers (but does not negate) the conversion rate of work into durable skill. Thus, a social planner interested in maximizing human capital has piece-rate preferences, as adolescent math proficiency is developed through experience in correctly solving a wide variety of problems, rather than by simply spending more time studying at home. This highlights how addressing productivity differentials may better enable low-performing students to achieve the work-volume needed for substantial skill gains.

The table also reports an average SD effect based on the *total derivative* of labor supply, or

$$E\left[\frac{d\Delta\kappa_i}{dA_i}\right]=E\left[\left(\frac{\partial\Delta\kappa_i}{\partial A_i}+\frac{\partial\Delta\kappa_i}{\partial T_i}\frac{\partial T_i}{\partial A_i}\right)\right],$$

which considers a change in work output, combined with the requisite time required by each student to complete that amount of extra work. The total derivative tends to be around 90% of the partial derivative with respect to A. A pSD increase in A implies an average improvement of 2.04(=0.681×2.991, specification (4)) exam-score points. Of course, these conversion rates are idiosyncratic and vary substantially across different children; Figure 8 (Appendix A) plots empirical CDFs of estimated partial derivatives, total derivatives, and net baseline learning.

6.3.3. *Skill Production Results: Age and Latent Student Traits.* While $(\theta_{pi}, \theta_{mi})$ are responsible for determining academic labor-supply (A_i, T_i), the primary inputs of skill growth, Table 6 considers whether UH traits also exert a direct influence on the *shape* of skill production technology. We find strong evidence that one's productivity type θ_p does so in a meaningful way, by advantageously influencing both slopes (p-value=0.015) and intercepts (p-value=0.093). Thus, students who are more productive with their learning time also tend to derive more durable skills from a fixed volume of completed homework as well. We find suggestive evidence that motivation θ_m has an advantageous direct impact on skill growth—entirely through slopes—but the estimated effect is somewhat noisy (p-value=0.170).

Holding productivity and academic work volume fixed, as children progress from 5$^{\text{th}}$-grade to 6$^{\text{th}}$-grade we find that they become more effective at learning. The total year-on-year difference in skill growth during our sample period—all else equal, through the age cohort dummy and interactions—is estimated at roughly 1/5 of a standard deviation. In other words, as students grow and gain more experience as learners, they not only augment their skill stock κ_i over time, but they also experience accelerated skill growth $\Delta\kappa_i$ as well.

6.3.4. *Dynamic Skill Acceleration Mechanisms.* So far, we have found evidence that learning accelerates over time, as the child gets older. This is in line with classic findings by Walberg and Tsai (1983), who coined the term, the "Matthew effect," to describe how early advantages in education tend to compound over time; Stanovich (1986), who showed "rich-get-richer" patterns in reading; and Aunola, Leskinen, Lerkkanen, and Nurmi (2004), who documented widening variance and faster growth among high-performing preschoolers. Our estimates point to similar dynamics in adolescent mathematics learning. This theme has been extensively studied in the economics literature on child development and human capital formation (e.g., Todd and Wolpin (2003); Cunha and Heckman (2007); Cunha et al. (2010); Heckman and Mosso (2014); Todd and Wolpin (2018)), but our combined empirical results in Sections 4, 5, and 6 prompt a reconsideration of the canonical "skills-beget-skills" insight.

Whereas traditional models emphasize *dynamic skill complementarity*—that is, early skill stocks raising the marginal productivity of future investments—our model of endogenous investment under multi-dimensional UH paints a somewhat more nuanced picture. Specifically, it highlights eight mechanisms affecting a child's rate of skill growth over time:

(1) ***Experience:*** $\varphi > 0$ (Section 4.4.1) $\Rightarrow$ investment A_i improves productivity, $\frac{d\rho_i(a)}{da} < 0$.
(2) ***Age-Productivity:*** Productivity improves ($\theta_{pi} \downarrow$) with age, beyond what can be explained solely by interim investment within the model (Section 5.1.3).
(3) ***Productivity-Investment:*** $\theta_{pi} \downarrow \Rightarrow (A_i, T_i) \uparrow$ (Section 2.5) $\Rightarrow \Delta\kappa_i \uparrow$ (Section 6.3.2).
(4) ***Motivation-Investment:*** $\theta_{mi} \downarrow \Rightarrow (A_i, T_i) \uparrow$ (Section 2.5) $\Rightarrow \Delta\kappa_i \uparrow$ (Section 6.3.2).
(5) ***Investment Efficiency:*** $\theta_{pi} \downarrow \Rightarrow \frac{T_i}{A_i} \downarrow$ (Section 2.1) $\Rightarrow \Delta\kappa_i \uparrow$ (Section 6.3.2).
(6) ***Productivity-Learning:*** $\theta_{pi} \downarrow \Rightarrow \Delta\kappa_i \uparrow$ (Section 6.3.2).
(7) ***Age-Learning:*** Older student $\Rightarrow \Delta\kappa_i \uparrow$ (Section 6.3.2).
(8) ***Skill Stock:*** Higher baseline skill $\kappa_i^{pre} \uparrow \Rightarrow \Delta\kappa_i \downarrow$ (Section 6.3.1).

Overall, these eight mechanisms point to *dynamic investment complementarity* as accelerating skill in various ways. The act of study and learning-by-doing is self-reinforcing, as experienced learners become more productive due to the feedback loop between mechanisms (1)–(4). Moreover, productivity improvements accelerate skill growth directly (mechanism (6)), and indirectly (mechanisms (3) and (5)). Growing older plays a significant advantageous role as well (mechanisms (2) and (7)), holding skill stock, UH, and labor-supply fixed. On the other hand, in adolescent mathematics, holding more prior skill stock has the effect of *decelerating* skill growth somewhat for the average student (mechanism (8)), though this effect is swamped by the other seven mechanisms.

6.4. **School VA Effects on Skill Production.** After controlling for a rich set of covariates, school quality plays an important role in skill growth. The VA ordering among the three school districts is consistent with previous stages of estimation: all else equal, a switch from District 1 to District 2 or 3 reduces average skill augmentation by 0.15 SD and 0.40 SD, resp., in specification (4). School district terms are highly significant all together, and individual District-1 and 2 total effects are still significant at the 1% level when taken separately.

The data support school quality impacts through both levels Δ_{0i} and slopes $\{\Delta_{1i}, \ldots, \Delta_{5i}\}$. The slope mechanism reflects schools' impact on the conversion rate of home study into math skill. The level mechanism may incorporate direct effects (e.g., more effective classroom teaching), and may also partly reflect impacts on the conversion rate of regular homework activity during the sample period. Overall, holding children's unobserved traits and external circumstances fixed, more effective schools act as a force multiplier for home study. Moreover, through the school-labor-supply interactions (slope terms) we can investigate school VA treatment heterogeneity as well. The data point to VA treatment effects varying widely: their standard deviations range between 69% and 138% of a SD of $\Delta\kappa$ in specification (4).

To sum up, our model allows for the following four channels through which better schools can create value added, and finds statistical evidence for three of them in our data:

(*i*) Augmenting academic productivity ($\downarrow \theta_{pi}$) (*significant*, Table 2);
(*ii*) Augmenting academic motivation ($\downarrow \theta_{mi}$) (insignificant, Table 3);
(*iii*) Augmenting skill growth beyond observed home study (*significant*, Table 6); and
(*iv*) Accelerating skill gains from home study (*significant*, Table 6)

6.4.1. *Threats to Causal Interpretation of School VA Effects.* School VA estimates shrink when controlling for UH in specification **(2)**, and remain stable across other specifications, which is suggestive of having adequately controlled for omitted variable bias. Since we were unable to include parent-survey controls in Table 6 (footnote 45), we perform a robustness check in Online Appendix D where we estimate a reduced-form model of initial math proficiency as a function of UH and the controls used in Table 6, plus parent survey variables. While this model does not have a structural interpretation as (12), it retains a similar school VA identification strategy. School VA results are similar and robust to parent-survey data.

The ability to address selection on unobservables by conditioning on UH has been central to identification in Section 6. To our knowledge, ours is the first paper to do so in estimating K-12 school VA.[46] The wealth of interactions between child UH, school dummies, and other variables also aid identification by reducing mis-specification bias. This is illustrated in Table 6: while the school VA average treatment effects are fairly stable across specifications, treatment-effect variance estimates change when the full set of controls is included. Of course, all claims to causality rest to some extent on unverifiable assumptions. We require that remaining unexplained skill growth $\log(\varepsilon)$ is conditionally uncorrelated with school dummies (Assumption 7). For school VA estimates to be largely attributable to omitted variable bias, there would have to exist some learning-relevant factor that is strongly correlated with school assignment, *and* not well proxied for by a combination of idiosyncratic study productivity and motivation, (θ_p, θ_m), plus our demographic, neighborhood, and household controls.

7. Conclusion

This paper advances the human-capital literature by uniting a structurally motivated field experiment with a model of academic labor supply in which both motivation and study productivity unobservably vary across students. We identify each dimension separately with real-time study data and randomized incentives, showing that a large share of observed achievement inequality reflects differences in productivity rather than in willingness to study. Conventional interventions that focus only on getting students to try harder risk overlooking the deeper and more prevalent underlying issues. Closing skill gaps in many cases requires policies that can raise the rate at which an hour of study is transformed into student success.

Our estimates underscore the analytical payoff to modeling multidimensional unobserved heterogeneity. Direct psychic costs dominate opportunity costs in shaping study decisions,

[46]Bodoh-Creed and Hickman (2024) estimate college VA by a similar strategy. Other work (e.g. Dale and Krueger (2002); Mountjoy and Hickman (2020)) indirectly controls for selection on unobservables in college VA using admissions portfolios.

and the marginal product of study time—correctly solved practice problems—has a more direct link to skill formation than raw time spent. Motivation and productivity interact non-linearly to produce choices and outcomes, and the same incentive schedule elicits very different labor-supply responses across students, holding either characteristic fixed. Interestingly, we find that productivity and motivation are uncorrelated. While productivity is a much stronger delineator of success versus struggle, on average, a high enough endowment of either trait may enable high achievement.

Our framework also provides a micro-foundation for school value-added. After conditioning on one's household, neighborhood, and UH traits, high-resource schools raise achievement chiefly by boosting students' home-study productivity and by strengthening the mapping from completed work to durable skill. This decomposition helps reconcile the modest test-score effects of many incentive or information programs with the much larger long-run impacts of attending an effective school documented elsewhere. Equity considerations loom large. Black and Hispanic students in our sample express higher interest in STEM studies, but tend to be less affluent and are mostly enrolled in low-resource schools with less advantageous value-added. Consequently, their average academic performance lags behind White and Asian students. The result is a painful irony: modern communities successfully market STEM to young people as a path out of poverty, yet fail to provide sufficient resources to underprivileged youth who are otherwise willing to tread that path.

There is, however, room for optimism. First, the evidence strongly suggests productivity is a malleable trait, responsive both to school quality and to learning experience. Second, we document highly motivated students among low-achieving groups, suggesting high returns to productivity-enhancing interventions targeted to such groups. Third, our evidence on decreasing returns-to-scale skill technology signals that the largest gains in aggregate math skill stock may arise from investments targeted toward under-served populations.

These findings recommend a shift in policy focus: from motivating students to making learning itself more productive, especially for those currently locked out of high-quality educational environments. In ongoing work, Cotton et al. (2025) take a deeper look into the framework presented here, by computing a series of policy counterfactuals that combine incentive-based interventions with productivity enhancements. Future work can advance our understanding by studying evolution of UH traits and endogenous academic effort over longer stretches of childhood, and by integrating additional dimensions of UH (e.g., attention, grit). Doing so will sharpen both positive and normative analyses of educational inequality.

Data Availability. Code and data for replicating tables and figures can be found in Cotton, Hickman, List, Price and Roy (2025), "Replication Data for: '*Why Don't Struggling Students Do Their Homework?: Disentangling Motivation and Study Productivity as Drivers of Human Capital Formation*'," in the Harvard Dataverse, https://doi.org/10.7910/DVN/OFYAMI.

Appendix A. Additional Figures

Figure 5. Website Task Completion by Gender and Race

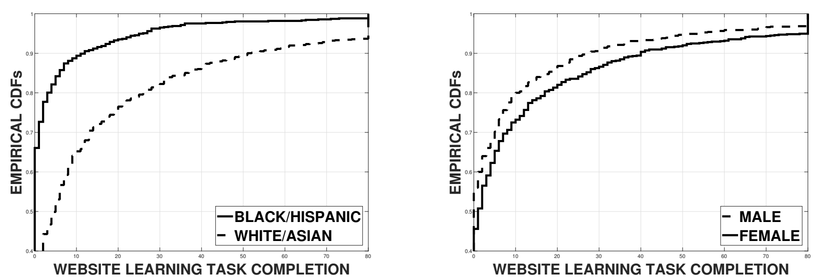


Figure 6. Cost Model Fit

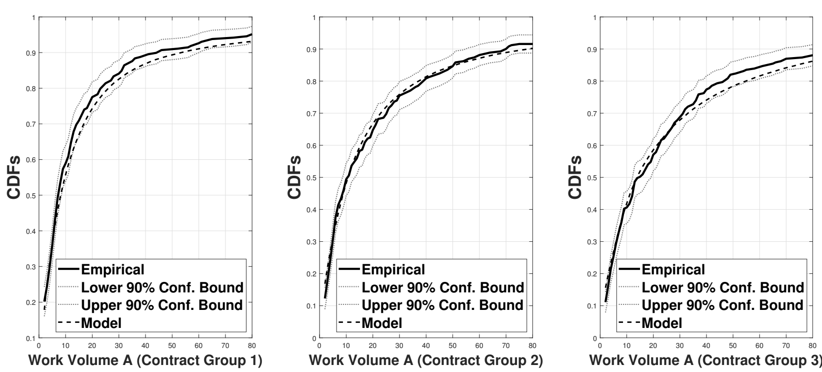


Figure 7. Empirical Bayes Shrunk Type Estimates

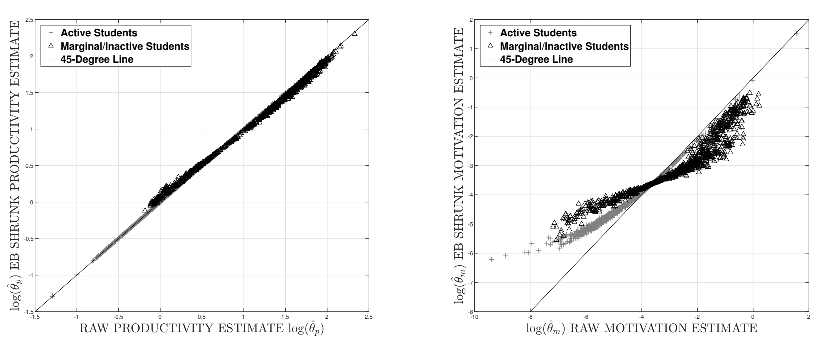
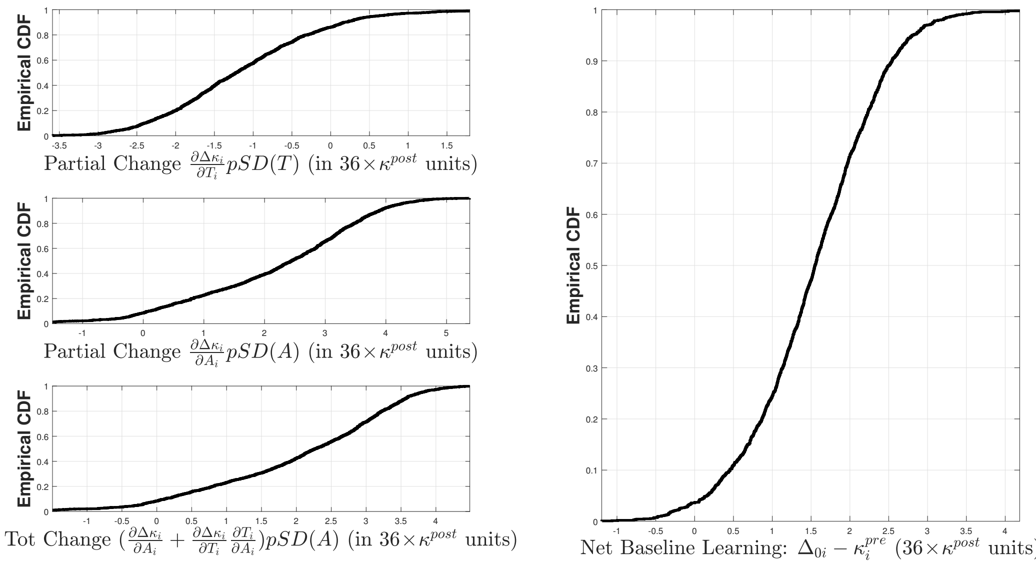


FIGURE 8. The Mechanics of Skill Acquisition



References

Abdulkadiroglu, A., Pathak, P. A., Schellenberg, J., & Walters, C. R. (2020). Do Parents Value School Effectiveness? *American Economic Review*, *110*, 1502–1539.

Agostinelli, F., Doepke, M., Sorrenti, G., & Zilibotti, F. (2025). It Takes a Village: The Economics of Parenting with Neighborhood and Peer Effects. *Journal of Political Economy*, *forthcoming*.

Agostinelli, F., Saharkhiz, M., & Wiswall, M. (2025). Home and School in the Development of Children. *Working paper*, *66(2)*, 527–566.

Agostinelli, F., & Wiswall, M. (2024). Estimating the Technology of Children's Skill Formation. *Journal of Political Economy*, *forthcoming*.

Ahn, T., Aucejo, E., & James, J. (2022). The Importance of Matching Effects for Labor Productivity: Evidence from Teacher-Student Interactions. *working paper, Arizona State University*.

Arrow, K. (1962). The Economic Implications of Learning by Doing. *The Review of Economic Studies*, *29(3)*, 155–173.

Arrow, K., Blackwell, D., & Girshick, A. (1949). Bayes and Minimax Solutions of Sequential Decision Problems. *Econometrica*, *17*, 213–244.

Attanasio, O., Cattan, S., Fitzsimmons, E., Meghir, C., & Rubio-Codina, M. (2020). Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia. *American Economic Review*, *110(1)*, 48–85.

Attanasio, O., Meghir, C., & Nix, E. (2020). Human Capital Development and Parental Investment in India. *Review of Economic Studies*, *87(6)*, 2511–2541.

Attanasio, O., & Pastorino, E. (2020). Nonlinear Pricing in Village Economies. *Econometrica*, *88(1)*, 207–263.

Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over Time: Dynamic Inconsistency in Real Effort Tasks. *Quarterly Journal of Economics*, *130(3)*, 1067–1115.

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental Dynamics of Math Performance From Preschool to Grade 2. *Journal of Educational Psychology*, *96*(4), 699–713.

Barrett, G., & Donald, S. (2003). Consistent Tests for Stochastic Dominance. *Econometrica*, *71*(1), 71–104.

Becker, G. S. (1993). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education,* 3rd *ed.* Chicago: University of Chicago Press.

Bénabou, R., & Tirole, J. (2003). Intrinsic and Extrinsic Motivation. *Review of Economic Studies*, *70*, 489–520.

Ben-Porath, Y. (1967). The Production of Human Capital and the Life Cycle of Earnings. *Journal of Political Economy*, *75(4)*, 352–365.

Bodoh-Creed, A., Hickman, B., & Boehnke, J. (2021). How Efficient are De-Centralized Auction Platforms? *Review of Economic Studies*, *88(1)*, 91–125.

Bodoh-Creed, A., Hickman, B., List, J., Muir, I., & Sun, G. (2024). Stress-Testing Structural Models of Unobserved Heterogeneity: Robust Inference on Optimal Nonlinear Pricing. *Working Paper, WashU-Olin*.

Bodoh-Creed, A., & Hickman, B. R. (2024). Pre-College Human Capital Investment and Affirmative Action: A Structural Policy Analysis of US College Admissions. *Working Paper, WashU-Olin*.

Brugués', F. (2024). Take the Goods and Run: Contracting Frictions and Market Power in Supply Chains. *working paper, Instituto Tecnológico Autónomo de México'*.

Burgess, S., Metcalfe, R., & Sadoff, S. (2016). *Understanding the Response to Financial and Non-Financial Incentives in Education: Field-Experimental Evidence Using High-Stakes Assessments* (No. 10284). (IZA Discussion Paper Series)

Carroll, J. B. (1963). A Model of School Learning. *Teachers College Record*, *64*, 723–733.

Chen, C., Jia, Y., Martin, X., & Silveira, B. (2024). Assessing the Objective Function of the SEC against Financial Misconduct: A Structural Approach. *forthcoming, Journal of Accounting and Economics*.

Chetty, R. (2012). Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply. *Econometrica*, *80(3)*, 969–1018.

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

Chetty, R., Hendren, N., & Katz, L. F. (2016). The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review*, *106*, 855–902.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2009). The Academic Achievement Gap in Grades 3 to 8. *Review of Economics and Statistics*, *91*, 398–419.

Cotton, C., Hickman, B., List, J., & Sun, G. (2025). How Can We Help Struggling Students Learn? Insights for Addressing Skill Gaps Under Adverse Selection in Human Capital. *WashU-Olin working paper*.

Cotton, C., Hickman, B. R., & Price, J. P. (2022). Affirmative Action and Human Capital Investment: Evidence from a Randomized Field Experiment. *Journal of Labor Economics*, *40(1)*, 157–185.

Cotton, C., Nordstrom, A., Nanowski, J., & Richert, E. (2024). Can Discussions about Girls' Education Improve Academic Outcomes? Evidence from a Randomized Development Project. *World Bank Economic Review*. (https://doi.org/10.1093/wber/lhae021)

Cullen, J., Levitt, S., Robertson, E., & Sadoff, S. (2013). The Academic Achievement Gap in Grades 3 to 8. *Journal of Economic Perspectives*, *27(2)*, 133–152.

Cunha, F., & Heckman, J. J. (2007). The Technology of Skill Formation. *AEA Papers & Proceedings*, *97*, 31–47.

Cunha, F., & Heckman, J. J. (2008). Formulating, identifying, and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, *43*, 738–782.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*, 883–931.

Cunha, F., Nielsen, E., & Williams, B. (2021). The Econometrics of Early Childhood Human Capital and Investments. *Annual Review of Economics*, *13*, 487–513.

Dale, S. B., & Krueger, A. B. (2002). Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables. *Quarterly Journal of Economics*, *117*, 1491–1528.

Del Boca, D., Flinn, C., Verriest, E., & Wiswall, M. (2024). Actors in the Child Development Process. *NBER Working Paper*, *#25596*.

Del Boca, D., Flinn, C., & Wiswall, M. (2014). Household Choices and Child Development. *Review of Economic Studies*, *81(11)*, 137–185.

Del Boca, D., Flinn, C., & Wiswall, M. (2016). Transfers to Households with Children and Child Development. *The Economic Journal*, *126*, 136–183.

DellaVigna, S., List, J., Malmendier, U., & Rao, G. (2022). Estimating Social Preferences and Gift Exchange at Work. *American Economic Review*, *112(3)*, 1038–1074.

D'Haultfoeuille, X., & Février, P. (2015). Identification of Triangular Nonseparable Models with Discrete Instruments. *Econometrica*, *83(3)*, 1199–1210.

D'Haultfoeuille, X., & Février, P. (2020). The Provision of Wage Incentives: A Structural Estimation Using Contracts Variation. *Quantitative Economics*, *11(1)*, 349–497.

Dobbie, W., & Fryer, R. (2011). Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, *3(3)*, 158–187.

Duckworth, A. L., Eichstaedt, J. C., & Ungar, L. H. (2007). The Mechanics of Human Achievement. *Social and Personality and Psychology Compass*, *9(7)*, 359–369.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Meece, C. M., & Midgley, C. (1983). Expectancies, Values, and Academic Behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives.* San Francisco: W. H. Freeman.

Fryer, R. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. *Quarterly Journal of Economics*, *126*, 1755–1798.

Fryer, R. (2016). Information, Non-Financial Incentives, and Student Achievement: Evidence from a Text Messaging Experiment. *Journal of Public Economics*, *144*, 109–121.

Fryer, R. (2017). Management and Student Achievement: Evidence from a Randomized Field Experiment. *NBER Working Paper*, *#23437*.

Fryer, R., Levitt, S., & List, J. (2015). Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights. *NBER Working Paper No. 21477*.

Fryer, R., Levitt, S., List, J., & Sadoff, S. (2022). Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment. *American Economic Journal: Economic Policy*, *14(4)*, 269–299.

Gayle, G.-L., Golan, L., & Soytas, M. (2022). What Accounts for the Racial Gap in Time Allocation and Intergenerational Transmission of Human Capital? *working paper, Washington University in St. Louis Economics Dept.*.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring Success in Education: The Role of Effort on the Test Itself. *American Economic Review: Insights*, *1*, 291–308.

Gneezy, U., & Rustichini, A. (2000). Pay Enough or Don't Pay at All. *Quarterly Journal of Economics*, *115*, 791–810.

Guerre, E., Perrigne, I., & Vuong, Q. (2009). Nonparametric Identification of Risk Aversion in First-Price Auctions Under Exclusion Restrictions. *Econometrica*, *77(4)*, 1193–1227.

Guryan, J., Ludwig, J., Bhatt, M., Cook, P., Davis, J., Dodge, K., ... Steinberg, L. (2021). Not Too Late: Improving Academic Outcomes Among Adolescents. *NBER Working Paper*, *#28531*.

Hanushek, E. A. (2020). Education production functions. In S. Bradley & C. Green (Eds.), *The Economics of Education: A Comprehensive Overview (2nd Ed.)* (pp. 161–170). Academic Press.

Hanushek, E. A., & Rivkin, S. G. (2006). School Quality and the Black-White Achievement Gap. *NBER Working Paper no 12651*.

Hanushek, E. A., & Rivkin, S. G. (2009). Harming the Best: How Schools Affect the Black-White Achievement Gap. *Journal of Policy Analysis and Management*, *28*, 366–393.

Heckman, J., & Mosso, S. (2014). The Economics of Human Development and Social Mobility. *Annual Review of Economics*, *6*, 689–733.

Heckman, J., & Zhou, J. (2025). A Study of the Microdynamics of Early Childhood Learning. *Journal of Political Economy*, *forthcoming*.

Hedblom, D., Hickman, B. R., & List, J. A. (2022). Toward and Understanding of Corporate Social Responsibility: Theory and Field Experimental Evidence. *Working Paper, WashU-Olin*.

Hirsch, W. Z. (1956). Firm Progress Ratios. *Econometrica*, *24(2)*, 136–143.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, *26(1)*, 101-136.

Judd, K. L. (1998). *Numerical Methods in Economics* (1st ed.). Cambridge, MA: The MIT Press.

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy.* Washington D.C.: Brookings Institution.

Kang, K., & Silveira, B. S. (2021). Understanding Disparities in Punishment: Regulator Preferences and Expertise. *Journal of Political Economy*, *129(10)*, 2947–2992.

Klenow, P. J. (1998). Learning Curves and the Cyclical Behavior of Manufacturing Industries. *Review of Economic Dynamics*, *1(2)*, 531–550.

Kleven, H., & Waseem, M. (2013). Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *The Quarterly Journal of Economics*, *128(2)*, 669–723.

Koedel, C., Mihaly, K., & Rockoff, J. (2015). Value-Added Modelling: A Review. *Economics of Education Review*, *47*, 180–195.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to Learn. *Review of Economics and Statistics*, *91*, 437–456.

Kurzban, R., Duckworth, A., Kable, J., & Meyers, J. (2013). An Opportunity Cost Model of Subjective Effort and Task Performance. *Behavioral and Brain Sciences*, *36(6)*, 661–726.

Leuven, E., Oosterbeek, H., & van der Klaauw, B. (2010). The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment. *Journal of the European Economic Association*, *8*, 1243–1265.

Levitt, S. D., List, J. A., & Sadoff, S. (2016). The Effect of Performance-Based Incentives on Educational Achievement: Evidence from a Randomized Experiment. *NBER Working Paper No. 22107*.

Little, R. J. A. (1992). Regression with Missing Xs: A Review. *Journal of the American Statistical Association*, *87*, 1227–1237.

Luccioni, M. (2024). The Determinants of Teaching Effectiveness: Evidence from a Model of Teachers' and Students' Interactions. *working paper,Olin Business School, Washington University in St Louis*.

Luo, Y., Perrigne, I., & Vuong, Q. (2018). Structural Analysis of Nonlinear Pricing. *Journal of Political Economy*, *126(6)*, 2523–2568.

Milgrom, P., & Shannon, C. (1994). Monotone Comparative Statics. *Econometrica*, *62(1)*, 157–180.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and application. *Journal of the American Statistical Association*, *78*, 47–55.

Mountjoy, J., & Hickman, B. R. (2020). The Return(s) to Colleges: Estimating Value-Added and Match Effects in Higher Education. *Becker-Friedman Institute Working Paper Series*, *2020-08*.

Rao, G. (2019). Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools. *American Economic Review*, *109(3)*, 774–809.

Rothstein, J. (2006). Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions. *American Economic Review*, *96(4)*, 1333–1350.

Saez, E. (2002). Optimal Transfer Programs: Intensive versus Extensive Labor Supply Responses. *The Quarterly Journal of Economics*, *117(3)*, 1039–1073.

Stanovich, K. E. (1986). Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *Reading Research Quarterly*, *22*(4), 360–407.

Sun, G. (2024). What, When, and Why Can the Principal Know: Identification in Adverse-Selection Models with an Application to Labor Income Taxation. *working paper, Washington University in St Louis Olin Business School*.

Thompson, P. (2010). Chapter 10 - Learning by Doing. In B. H. Hall & N. Rosenberg (Eds.), *Handbook of the economics of innovation, vol. 1* (Vol. 1, p. 429-476). North-Holland.

Todd, P., & Wolpin, K. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, *113*, 3–33.

Todd, P., & Wolpin, K. (2018). Accounting for Mathematics Performance of High School Students in Mexico: Estimating a Coordination Game in the Classroom. *Journal of Political Economy*, *126(6)*, 2608–2650.

Torgovitsky, A. (2015). Identification of Nonseparable Models Using Instruments with Small Support. *Econometrica*, *83(3)*, 1185–1197.

Verbeek, M., & Nijman, T. (1992). Testing for Selectivity Bias in Panel Data Models. *International Economic Review*, *33(3)*, 681–703.

Walberg, H. J., & Tsai, S.-L. (1983). Matthew Effects in Education. *American Educational Research Journal*, *20*(3), 359–373.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, *16(2)*, 117–186.

Williams, B. (2019a). Controlling for Ability Using Test Scores. *Journal of Applied Econometrics*, *34*(4), 547–565.

Williams, B. (2019b). Identification of a Nonseparable Model Under Endogeneity Using Binary Proxies for Unobserved Heterogeneity. *Quantitative Economics*, *10*(2), 527–563.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach, 6th Edition*. Boston, MA: Cenage Learning.

Wright, T. P. (1936). Factors Affecting the Cost of Airplanes. *Journal of the Aeronautical Sciences*, *3(4)*, 122–128.

Table 1. WEBSITE MATH ACTIVITY & DAILY HOMEWORK TIME

Variable	**Sample Mean**	**Sample Median**	**Sample Std. Dev.**	**N**	**Contract Group 1 Mean**	**Contract Group 2 Mean**	**Contract Group 3 Mean**
MASSES AT DIFFERENT WEBSITE ACTIVITY LEVELS							
Active Students $\mathbb{1}(\boldsymbol{A_i \geq 2})$	0.447	0	0.497	1,676	0.422	0.453	0.466
Marginal Students $\mathbb{1}(\boldsymbol{A_i = 1})$	0.056	0	0.230	1,676	0.072	0.043	0.054
Inactive Students $\mathbb{1}(\boldsymbol{A_i = 0})$	0.497	0	0.500	1,676	0.506	0.504	0.480
EXTRACURRICULAR MATH ACTIVITY, CONDITIONAL ON $A_i \geq 2$							
Learning Tasks Completed $(\boldsymbol{A_i})$	22.35	12	24.29	749	17.72	22.91	25.98
Math Problems Solved	134.11	72	145.75	749	106.31	137.48	155.89
Website Time (min.) $(\boldsymbol{T_i})$	157.05	102.85	152.45	749	122.74	160.13	184.96
Within-Child Avg. Time Per Comp. Task (min.) $\left(\frac{\boldsymbol{T_i}}{\boldsymbol{A_i}}\right)$	10.33	7.84	7.38	749	—	—	—
Within-Child Computer Pageload Fraction	0.768	1	0.376	749	—	—	—
Within-Child Tablet Pageload Fraction	0.185	0	0.348	749	—	—	—
Within-Child Smartphone Pageload Fraction	0.048	0	0.172	749	—	—	—
Total Pay	\$33.05	\$21.75	\$25.77	749	\$28.29	\$32.91	\$37.48
Avg. Piece-Rate Wage/Hr	\$8.52	\$7.42	\$5.45	749	\$6.37	\$8.39	\$10.59
SELF-REPORTED AVG. DAILY HOMEWORK TIME ACROSS ALL ACADEMIC SUBJECTS							
All Students (hrs)	1.248	1.214	0.681	1,676	—	—	—
Active Only (hrs)	1.422	1.429	0.646	749	—	—	—
(95% Conf. Int.)	*(1.38,1.47)*						
Marg./Inactive (hrs)	1.108	1.071	0.677	927	—	—	—
(95% Conf. Int.)	*(1.06,1.15)*						

Table 2. TOBIT REGRESSION RESULTS: PRODUCTIVITY

DEP VAR:	(1)		(2)		(3)		(4)		(5)	
log $(\boldsymbol{\theta_p})$	Estimate	SD Effect	Estimate	SD Effect	Estimate	SD Effect	Estimate	SD Effect	Estimate	SD Effect
Female	0.233***	**0.240**	0.222***	**0.229**	0.132***	**0.158**	0.170***	**0.211**	0.233***	**0.303**
(std. err.)	*(0.012)*		*(0.009)*		*(0.017)*		*(0.036)*		*(0.032)*	
Black	0.894***	**0.917**	0.721***	**0.742**	0.817***	**0.982**	0.767***	**0.950**	0.753***	**0.979**
(std. err.)	*(0.018)*		*(0.015)*		*(0.040)*		*(0.073)*		*(0.040)*	
Hispanic	0.805***	**0.826**	0.612***	**0.630**	0.517***	**0.622**	0.560***	**0.694**	0.525***	**0.683**
(std. err.)	*(0.019)*		*(0.023)*		*(0.041)*		*(0.080)*		*(0.054)*	
Grade 5	0.272***	**0.279**	0.290***	**0.299**	0.269***	**0.324**	0.268***	**0.332**	0.233***	**0.302**
(std. err.)	*(0.010)*		*(0.007)*		*(0.014)*		*(0.025)*		*(0.024)*	
Constant	0.064***	—	-	—	-	—	-	—	-	—
			0.067***		0.261***		0.310***		0.500***	
(std. err.)	*(0.017)*		*(0.019)*		*(0.027)*		*(0.069)*		*(0.032)*	
SCHOOL EFFECTS (District 1 omitted)										
Joint Exclusion P-Value $(df=2)$:			$<10^{-16}$		$<10^{-16}$		$2.2{\times}10^{-14}$		$<10^{-16}$	
District 2	—	—	0.206***	**0.212**	0.150***	**0.180**	0.153***	**0.190**	0.091**	**0.118**
(std. err.)			*(0.009)*		*(0.027)*		*(0.055)*		*(0.036)*	
District 3	—	—	0.610***	**0.628**	0.577***	**0.694**	0.503***	**0.624**	0.500***	**0.650**
(std. err.)			*(0.018)*		*(0.044)*		*(0.066)*		*(0.061)*	
Nbhd-SES Cntrls (2)	YES***		YES		YES		YES		YES***	
Home Acad Support (2)	no		no		YES***		YES**		YES***	
Home Connectivity (3)	no		no		no		YES***		YES***	
Prnt Survey Cntrls (4)	no		no		no		no		YES***	
$\boldsymbol{N}$	1,676		1,676		1,676		1,676		1,676	
log-$\boldsymbol{\mathcal{L}}$	-3,455.7		-3,444.5		-3,358.7		-3,336.0		-3,310.0	
Pseudo-$\boldsymbol{R^2}$ $(\boldsymbol{log(\theta_p)})$	0.356		0.349		0.419		0.442		0.437	

Notes: **SD Effect** is the change in standard deviation units of $\log(\theta_p)$ of switching a binary regressor value from 0 to 1; bold font indicates significance a the 90% level or higher. Significance of coefficient estimates at the 90%, 95%, and 99% levels are denoted by "$*$," "$**$," and "$***$," respectively. Stars on YES/no entries indicate joint statistical significance level for all variables within that group (via a Wald test). **Pseudo-$\boldsymbol{R^2}$** is $1 - \mathrm{E}[Var(\hat{\eta}_{pi})] / \Big(Var(\boldsymbol{X}_{pi}\hat{\beta}_p) + \mathrm{E}[Var(\hat{\eta}_{pi})]\Big)$.

Table 3. TOBIT REGRESSION RESULTS: MOTIVATION

DEP VAR:	**(1)**		**(2)**		**(3)**		**(4)**		**(5)**	
$\log(\theta_m)$	Estimate	SD Effect	Estimate	SD Effect	Estimate	SD Effect	Estimate	SD Effect	Estimate	SD Effect
Female	-1.304***	**-0.781**	-1.275***	**-0.774**	-0.800***	**-0.390**	-0.770	-0.379	-0.861***	**-0.425**
(std. err.)	*(0.085)*		*(0.102)*		*(0.230)*		*(0.810)*		*(0.287)*	
Black	-0.907***	**-0.543**	-0.883***	**-0.536**	-1.720*	**-0.838**	-1.336	-0.658	-1.274	-0.629
(std. err.)	*(0.166)*		*(0.216)*		*(0.929)*		*(0.925)*		*(0.842)*	
Hispanic	-0.093	-0.056	-0.101	-0.061	-0.296	-0.144	-0.460	-0.227	-0.376	-0.186
(std. err.)	*(0.242)*		*(0.874)*		*(0.953)*		*(0.988)*		*(1.143)*	
Grade 5	-0.145**	**-0.087**	-0.161**	**-0.098**	-0.079	-0.038	-0.152	-0.075	-0.047	-0.023
(std. err.)	*(0.070)*		*(0.074)*		*(0.226)*		*(0.753)*		*(0.345)*	
Math Favorite	—	—	—	—	-0.316	-0.154	-0.348	-0.172	-0.332	-0.164
(std. err.)					*(0.221)*		*(0.835)*		*(0.301)*	
Math Least Favorite	—	—	—	—	0.326	0.159	0.299	0.147	0.318	0.157
(std. err.)					*(0.434)*		*(0.885)*		*(0.411)*	
Intrinsic Score	—	—	—	—	-0.884***	**-0.431**	-0.815*	**-0.360**	-0.833***	**-0.369**
(std. err.)					*(0.196)*		*(0.476)*		*(0.192)*	
Extrinsic Score	—	—	—	—	-0.959***	**-0.468**	-1.001**	**-0.413**	-0.984***	**-0.407**
(std. err.)					*(0.263)*		*(0.451)*		*(0.215)*	
Reg Study Time	—	—	—	—	—	—	-0.210	-0.070	-0.177	-0.059
(std. err.)							*(0.529)*		*(0.278)*	
Recr Screen Time	—	—	—	—	—	—	0.193	0.116	0.180	0.108
(std. err.)							*(0.180)*		*(0.181)*	
Constant	-4.102***	—	-4.111***	—	-1.482*	—	-1.508	—	-1.426	—
(std. err.)	*(0.089)*		*(0.278)*		*(0.847)*		*(1.058)*		*(0.998)*	
SCHOOL EFFECTS (District 1 omitted)										
Joint Exclusion P-Value ($df = 2$):			0.984		0.929		0.976		0.892	
District 2	—	—	0.018	0.011	0.275	0.134	0.157	0.077	0.273	0.135
(std. err.)			*(0.155)*		*(0.724)*		*(0.870)*		*(0.604)*	
District 3	—	—	-0.002	-0.001	0.018	0.009	0.135	0.066	0.177	0.087
(std. err.)			*(0.350)*		*(0.908)*		*(0.978)*		*(1.187)*	
Nbhd-SES Cntrls (2)	YES		YES		YES		YES		YES	
Home Acad Support (2)	no		no		YES		YES		YES	
Extracurriculars (4)	no		no		no		YES		YES	
Gaming/Surfing (3)	no		no		no		YES		YES	
Home Connectivity (3)	no		no		no		YES		YES	
Parent Srvy Cntrls (4)	no		no		no		no		YES	
N	1,676		1,676		1,676		1,676		1,676	
log-$\mathcal{L}$	-3,455.7		-3,444.5		-3,358.7		-3,336.0		-3,310.0	
Pseudo-R^2 ($\log(\theta_m)$)	0.207		0.202		0.302		0.298		0.293	

Notes: **SD Effect** is the change in standard deviations of $\log(\theta_m)$ of switching a binary regressor value from 0 to 1, or from increasing a continuous regressor value by one standard deviation; bold font indicates significance at the 90% level or higher. Significance of coefficient estimates at the 90%, 95%, and 99% levels are denoted by "$*$," "$**$," and "$***$," respectively. Stars on YES/no entries indicate joint statistical significance level for all variables within that group (via a Wald test). **Pseudo-R^2** is $1 - \mathrm{E}[Var(\hat{\eta}_{mi})] / \left(Var(\boldsymbol{X}_{mi}\hat{\beta}_m) + \mathrm{E}[Var(\hat{\eta}_{mi})]\right)$.

Table 4. TOBIT STANDARD DEVIATIONS AND CORRELATIONS

	TOBIT SPECIFICATION (4)						**TOBIT SPECIFICATION (5)**					
	$\boldsymbol{X}_{pi}\boldsymbol{\beta}_p$	$\boldsymbol{X}_{mi}\boldsymbol{\beta}_m$	η_{pi}	η_{mi}	$\log(\theta_{pi})$	$\log(\theta_{mi})$	$\boldsymbol{X}_{pi}\boldsymbol{\beta}_p$	$\boldsymbol{X}_{mi}\boldsymbol{\beta}_m$	η_{pi}	η_{mi}	$\log(\theta_{pi})$	$\log(\theta_{mi})$
$\widehat{\boldsymbol{StDev}}$	0.532	1.108	0.598	1.708	0.800	2.036	0.509	1.096	0.569	1.710	0.763	2.031
(StdErr)	*(0.020)*	*(0.306)*	*(0.025)*	*(0.170)*	*(0.019)*	*(0.289)*	*(0.013)*	*(0.178)*	*(0.018)*	*(0.139)*	*(0.011)*	*(0.177)*
$\widehat{\boldsymbol{Correl.}}$	-0.171		-0.082		-0.113		-0.077		-0.033		-0.048	
(StdErr)	*(0.123)*		*(0.184)*		*(0.131)*		*(0.130)*		*(0.129)*		*(0.115)*	
95% CI	[-0.39,0.05]		[-0.45,0.27]		[-0.37,0.13]		[-0.32,0.16]		[-0.28,0.19]		[-0.27,0.15]	
P-value	0.164		0.657		0.388		0.554		0.800		0.675	

Table 5. DESCRIPTIVE STATISTICS: EXAM SCORE CHANGES BY SUB-SAMPLE

SUB-SAMPLE:	**ALL**	**FEMALE**	**MALE**	**BLACK**	**HISPANIC**	**WHITE/ASIAN**
SIZE/FRACTION:	1,494	0.516	0.483	0.254	0.187	0.560
Pre-Test Score, $36\times\boldsymbol{S^{pre}}$**:**	16.492	15.62	17.42	11.56	11.70	20.33
(Sample StDev)	*(8.042)*	*(7.39)*	*(8.60)*	*(5.65)*	*(5.70)*	*(7.55)*
($36\times\kappa^{pre}$ StDev, EB)	*(5.001)*	—	—	—	—	—
Post-Test Score, $36\times\boldsymbol{S^{post}}$**:**	18.079	17.59	18.61	12.52	12.36	22.51
(Sample StDev)	*(8.998)*	*(8.58)*	*(9.41)*	*(6.82)*	*(6.13)*	*(8.16)*
($36\times\kappa^{post}$ StDev, EB)	*(5.709)*	—	—	—	—	—
$36\times\boldsymbol{\Delta S}$ **(All):**	1.587	1.97	1.18	0.96	0.66	2.18
(Sample StDev)	*(4.653)*	*(4.74)*	*(4.53)*	*(4.70)*	*(4.61)*	*(4.56)*
($36\times\Delta\kappa$ StDev, EB)	*(2.991)*	—	—	—	—	—
(p-value, $H_0:E[\Delta S]=0$)	1.3×10^{-37}	1.8×10^{-28}	5.2×10^{-12}	8.8×10^{-5}	0.018	2.3×10^{-39}
$36\times\boldsymbol{\Delta S}$ **(Active):**	2.603	$36\times\boldsymbol{\Delta S}$ **(Marg./Inact.):**		0.649		
(Sample StDev)	*(4.571)*	*(Sample StDev)*		*(4.533)*		
(p-value, $H_0:E[\Delta S]=0$)	1.1×10^{-45}	*(p-value, $H_0:E[\Delta S]=0$)*		7.1×10^{-5}		

Notes: S^{pre} and S^{post} are measured as the fraction of math problems correctly solved on a 36-problem test, with a 35-minute time limit. Sample means are in regular font; sample standard deviations are italicized. Note that $E[S^{pre}]=\kappa^{pre}$, and "EB" denotes the empirical Bayes StDev estimate. A two-sample t-test rejects the null hypothesis that website activity did not impact learning, $H_0:E[\Delta S|Active]=E[\Delta S|Marg./Inactive]$ (p-value$=2.5\times10^{-16}$). 5^{th}-graders are 47.3% of the sample.

Table 6. PRODUCTION TECHNOLOGY OF MATH SKILL GAINS

SPECIFICATION: **DEP VAR:** $\boldsymbol{36\times S_{EBi}^{post}}$	**(1)** **(Mean; SD)**	**(2)** **(Mean; SD)**	**(3)** **(Mean; SD)**	**(4)** **(Mean; SD)**
SKILL GROWTH: $\boldsymbol{36\times\Delta\kappa_i}$	(1.587; 2.991)	(1.587; 2.991)	(1.587; 2.991)	(1.587; 2.991)
Net Baseline Learning $\boldsymbol{\hat{\Delta}_{0i}-36\times\kappa_i^{pre}}$	(1.419; 0.647)	(1.671; 0.722)	(1.516; 0.799)	(1.535; 0.821)
	Avg SD Effect	**Avg SD Effect**	**Avg SD Effect**	**Avg SD Effect**
PRIOR SKILL STOCK EFFECTS:				
$\boldsymbol{36\times\kappa_i^{pre}}$	-0.336***	-0.369***	-0.395***	-0.418***
joint p-value: $\hat{\delta}_{01},\ldots,\hat{\delta}_{51}$	*($<10^{-16}$)*	*($<10^{-16}$)*	*($<10^{-16}$)*	*($<10^{-16}$)*
$\boldsymbol{36\times\kappa_i^{pre}}$**, Slopes Only**	-0.098***	-0.086***	-0.072***	-0.088***
joint p-value: $\hat{\delta}_{11},\ldots,\hat{\delta}_{51}$	*($<10^{-16}$)*	*(0.008)*	*(0.005)*	*(0.009)*
$\boldsymbol{36\times\kappa_i^{pre}}$**, Intercept Term Only**	4.277***	4.141***	4.020***	4.002***
p-value: $\hat{\delta}_{01}$	*($<10^{-16}$)*	*($<10^{-16}$)*	*($<10^{-16}$)*	*($<10^{-16}$)*
HOME STUDY EFFECTS:				
Partial Change $\boldsymbol{E\left[\frac{\partial\Delta\kappa_i}{\partial T_i}\right]\frac{pSD(T)}{S(\Delta\kappa)}}$	-0.797***	-0.309***	-0.182***	-0.379***
joint p-value: $\hat{\Delta}_{1i},\hat{\Delta}_{2i},\hat{\Delta}_{5i}$	*($<10^{-16}$)*	*($<10^{-16}$)*	*($<10^{-16}$)*	*($<10^{-16}$)*
Partial Change $\boldsymbol{E\left[\frac{\partial\Delta\kappa_i}{\partial A_i}\right]\frac{pSD(A)}{S(\Delta\kappa)}}$	1.465***	0.539***	0.396***	0.729***
joint p-value: $\hat{\Delta}_{3i},\hat{\Delta}_{4i},\hat{\Delta}_{5i}$	*($<10^{-16}$)*	*(4.3×10^{-11})*	*(1.5×10^{-14})*	*($<10^{-16}$)*
Total Change $\boldsymbol{E\left[\frac{d\Delta\kappa_i}{dA_i}\right]\frac{pSD(A)}{S(\Delta\kappa)}}$	**1.377**	**0.493**	**0.361**	**0.681**
STUDENT TRAIT EFFECTS:				
Inv. Prod. log($\boldsymbol{\theta_{pi}}$ **)**	—	-0.371***	-0.391**	-0.362***
joint p-value: $\hat{\delta}_{02},\ldots,\hat{\delta}_{52}$		*(0.003)*	*(0.002)*	*(0.002)*
Inv. Prod. log($\boldsymbol{\theta_{pi}}$ **), Slopes Only**	—	-0.149	-0.246**	-0.196**
joint p-value: $\hat{\delta}_{12},\ldots,\hat{\delta}_{52}$		*(0.130)*	*(0.029)*	*(0.015)*
Inv. Motiv. log($\boldsymbol{\theta_{mi}}$ **)**	—	-0.003	-0.028	-0.071
joint p-value: $\hat{\delta}_{03},\ldots,\hat{\delta}_{53}$		*(0.758)*	*(0.669)*	*(0.236)*
Inv. Motiv. log($\boldsymbol{\theta_{mi}}$ **), Slopes Only**	—	0.030	-0.032	-0.075
joint p-value: $\hat{\delta}_{13},\ldots,\hat{\delta}_{53}$		*(0.641)*	*(0.543)*	*(0.170)*
Grade 5	—	—	-0.206***	-0.208***
joint p-value: $\hat{\delta}_{04},\ldots,\hat{\delta}_{54}$			*(1.7×10^{-4})*	*(1.0×10^{-6})*
SCHOOL EFFECTS (DISTRICT 1 OMITTED):				
Joint P-Value, All Terms (df = 12):	*($<10^{-16}$)*	*(1.6×10^{-15})*	*($<10^{-16}$)*	*(1.8×10^{-5})*
District 2	-0.270***	-0.147***	-0.240***	-0.154***
joint p-value: $\hat{\delta}_{08},\ldots,\hat{\delta}_{58}$	*(2.5×10^{-5})*	*(0.004)*	*($<10^{-16}$)*	*(8.2×10^{-6})*
District 2, Slopes Only	0.028**	-0.068**	-0.055***	-0.028***
joint p-value: $\hat{\delta}_{18},\ldots,\hat{\delta}_{58}$	*(0.020)*	*(0.041)*	*(5.0×10^{-13})*	*(2.3×10^{-5})*
Dist 2 Trtmt-Effect StdDev ($\Delta\kappa$ units):	*0.421*	*0.733*	*0.885*	*2.049*
District 3	-0.558***	-0.374***	-0.484***	-0.404***
joint p-value: $\hat{\delta}_{09},\ldots,\hat{\delta}_{59}$	*($<10^{-16}$)*	*(3.1×10^{-16})*	*($<10^{-16}$)*	*(2.6×10^{-4})*
District 3, Slopes Only	-0.072***	-0.152***	-0.064***	-0.061***
joint p-value: $\hat{\delta}_{19},\ldots,\hat{\delta}_{59}$	*(7.9×10^{-11})*	*(2.0×10^{-7})*	*(4.3×10^{-9})*	*(0.007)*
Dist 3 Trtmt-Effect StdDev ($\Delta\kappa$ units):	*1.072*	*1.148*	*1.510*	*4.138*
Race/Gender (18)	no	no	YES***	YES***
Nbhd-SES Cntrls (12)	no	no	no	YES***
Home Acad Support (6)	no	no	no	YES*
Home Connectivity (18)	no	no	no	YES***
$\boldsymbol{N=1{,}494}$**; (#Parameters,** $\boldsymbol{R^2}$**):**	(24, 0.781)	(36, 0.785)	(60, 0.791)	(96, 0.795)

Notes: **Avg SD Effect** for continuous variable z is defined as $\frac{\partial\Delta\kappa}{\partial z}\frac{SD(Z)}{SD(\Delta\kappa)}$, and for binary variables the derivative is replaced with a difference. To convert average SD effects to SD units of skill κ^{post} (rather than skill growth units), simply multiply by $\frac{2.991}{5.709}=0.524$. **Total Change** is $E\left[\frac{d\Delta\kappa_i}{dA_i}\right]\frac{pSD(A)}{SD(\Delta\kappa)}=E\left[\left(\frac{\partial\Delta\kappa_i}{\partial A_i}+\frac{\partial\Delta\kappa_i}{\partial T_i}\frac{\partial T_i}{\partial A_i}\right)\right]\frac{pSD(A)}{SD(\Delta\kappa)}$, where $\frac{\partial T_i}{\partial A_i}\approx\mathcal{T}_0\theta_{pi}A_i^{-\varphi}$ is used as a within-i median approximation. Significance at the 99%, 95% and 90% levels are denoted by ***, **, and *, respectively.

Table 7. COVARIATE MEANS & STD DEVIATIONS

SAMP SIZE: 1,676	Mean	(*StDev,* [frac. missing])		Mean	(*StDev,* [frac. missing])
SCHOOL DISTRICT & NEIGHBORHOOD SOCIOECONOMICS					
Nbhd Avg Income ($1,000's)	$101.7	(*41.5,* [0.05])	District 1	0.465	(*0.499,* [0.00])
Nbhd Uninsured Minors	0.252	(*0.297,* [0.06])	District 2	0.268	(*0.443,* [0.00])
			District 3	0.267	(*0.442,* [0.00])
FAMILY & RECREATIONAL TIME-USE VARIABLES					
#Adult Academic Helpers	1.140	(*0.848,* [0.07])	Enrolled in Sports	0.669	(*0.471,* [0.00])
#Peer Academic Helpers	0.789	(*0.783,* [0.07])	Enrolled in Music	0.383	(*0.487,* [0.00])
#Gaming Systems at Home	1.570	(*1.135,* [0.00])	Enrolled in Clubs/ Other Activities	0.410	(*0.492,* [0.00])
Parental Permission for Video Gaming on Weekdays	0.878	(*0.327,* [0.00])	Fraction Leisure Time In Adult-Supervised Activity	0.351	(*0.172,* [0.05])
Weekday Daily Recreational Internet Use (hrs)	1.766	(*1.201,* [0.09])			
ACADEMIC PREFERENCES & ATTITUDE VARIABLES					
Math Favorite Subj.	0.361	(*0.480,* [0.07])	Extrinsic Motiv. Score	0	(*1,* [0.00])
Math Least Favorite Subj.	0.216	(*0.411,* [0.08])	Intrinsic Motiv. Score	0	(*1,* [0.00])
DEMOGRAPHICS & DEMOGRAPHIC CORRELATIONS					
Black	0.269	(*0.444,* [0.00])	Female	0.508	(*0.500,* [0.00])
Hispanic	0.192	(*0.394,* [0.00])	Male	0.492	(*0.500,* [0.00])
White/Asian/Other	0.539	(*0.499,* [0.00])	Grade 5	0.473	(*0.499,* [0.00])
Nbhd Avg Inc (Blk,Hsp,WAO)	($77.5, $54.0, $129.6)		District 1 (Blk,Hsp,WAO)	(0.004, 0.018, 0.978)	
Math Favorite (Blk,Hsp,WAO)	(0.431, 0.439, 0.302)		District 2 (Blk,Hsp,WAO)	(0.653, 0.073, 0.274)	
STEM Favorite (Blk,Hsp,WAO)	(0.719, 0.733, 0.443)		District 3 (Blk,Hsp,WAO)	(0.346, 0.612, 0.042)	

Notes: un-bracketed, standard font numbers are sample means; italicized numbers are sample standard deviations; and bracketed, standard font numbers report fractions of missing observations. **Adult Academic Helpers** included parents, grandparents, and tutors; **Peer Academic Helpers** included siblings and friends. **Extrinsic Motivation Score** and **Intrinsic Motivation Score** both exist on a scale of 0-4, but have been standardized for this table. Fifth-graders make up 47.3% of the total sample, with 6^{th}graders comprising the other 52.7%. Sub-sample proportions are close to that ratio for all gender and race groups.

Table 8. DESCRIPTIVE STATISTICS: PARENT SURVEY VARIABLES

	N Obs.	%Sampled	Mean	Median	StDev	Min	Max
Involved Parent	333	19.87%	0.426	0	0.495	0	1
Big Family	307	18.32%	0.435	0	0.497	0	1
Youngest Child	307	18.32%	0.316	0	0.466	0	1
Middle Child	307	18.32%	0.502	1	0.501	0	1

Notes: **Involved Parent** is a dummy for (self-reported) daily average time spent with child on schoolwork weakly exceeding 2 hours. **Big Family** is a dummy for 3 or more children living in the household. **Youngest Child** (i.e., at least one older and no younger siblings) and **Middle Child** (i.e., at least one younger and one older sibling) are birth-order dummies; the omitted category is **Oldest Child**, which includes only-child status as a special case.

Table 9. COMMON STRUCTURAL PARAMETERS

KNOT LOCATIONS: (quoted in units of minute spent over a 10-day sample period) $\{\kappa_{c1}, \kappa_{c2}, \kappa_{c3}, \kappa_{c4}, \kappa_{c5}, \kappa_{c6}, \kappa_{c7}, \kappa_{c8}\} = \{0,\ 28.02,\ 46.12,\ 75.33,\ 109.59,\ 171.31,\ 289.31,\ 1254\}$

VARIABLE:	γ_{c1}	γ_{c2}	γ_{c3}	γ_{c4}	γ_{c5}	γ_{c6}
Point Est.:	0	9.3340	24.720	147.59	424.56	931.45
90% CI:	—	—	[24.19, 25.27]	[134.3, 157.9]	[403.4, 456.8]	[887.7, 981.6]
VARIABLE:	γ_{c7}	γ_{c8}	γ_{c9}	γ_{c10}		
Point Est.:	1,936.4	9,969.8	17,626	85,103		
90% CI:	[1760.5, 2100.3]	[9044.4, 11063]	[15246, 20532]	[72252, 111460]		
VARIABLE:	$\mathcal{T}_0$	$\mathcal{T}_1$	φ			
Point Est.:	6.575	1.058	0.0788			
90% CI:	[6.483, 6.648]	[1.052, 1.064]	[0.0741, 0.0817]			

Notes: Confidence intervals are computed via the bootstrap (400 bootstrap samples). γ_{c1} and γ_{c2} are not free parameters during estimation (pinned down by the boundary conditions $c(0 = 0$ and $c'(0) = 1$).

Online Supplement to Accompany

Why Don't Struggling Students Do Their Homework? Disentangling Motivation and Study Productivity as Drivers of Human Capital Formation,

by C. Cotton, B. Hickman, J. List, J. Price, and S. Roy

Appendix B. Additional Field Experiment Details

B.1. **Ethical Vetting of the Field Experiment.** Experimental procedures underwent ethical vetting by multiple IRBs (UChicago, UMiami, and BYU). School administrators and teachers from the three districts cooperated with the research team for this study, and served as the primary interface with student test subjects. The research team prepared all relevant research materials, which were distributed and collected to/from students by their math teachers. Participation was on an opt-out basis, meaning that (after prior notification) students in each math class were included in the study unless the child or his/her parent declined. Prior to the study, a parental assent form was emailed to parents, and hard copies went home with students. This form described the study, gave contact information for the research team, and described data-security measures it would follow. The assent form allowed parents to opt their child out of the study by signing and returning it, or by responding via email. On the first day of the study, students received an additional child consent form with similar information stated in age-appropriate language. This form emphasized that participation was optional and would not affect their academic standing. It gave each child an opportunity to opt out of their own volition. Language on both assent forms was scrutinized by three research ethics boards. Parents and students received multiple notifications—before *and* after data collection—of their right to withdraw from the study.

All study participants were made aware that the website used for the field experiment (www.ChigagoMathGame.Org) was hosted by the research team; e.g., the login page bore the logos of UChicago, UMiami, and BYU. They were also made aware that the website was tracking their activity for the purpose of computing their earnings during the study, and for the purpose of scientific research afterward. The opt-out participation setup carefully balanced scientific needs (a large, representative sample of the local student population), with ethical imperatives of clearly articulating study procedures and community members' rights, and providing ample opportunity to decline participation. A small fraction of students were opted out ($< 5\%$), but teachers and parents generally welcomed our study enthusiastically as a supplemental learning opportunity for their students. While data analyses focus solely on children in non-special-needs classes, some parents of special-needs students contacted us to request website/incentive participation by their child; we were happy to oblige.

All data publicly released in the replication files—Cotton, Hickman, List, Price and Roy (2025), "Replication Data for: '*Why Don't Struggling Students Do Their Homework?: Disentangling Motivation and Study Productivity as Drivers of Human Capital Formation*'," in the Harvard Dataverse, https://doi.org/10.7910/DVN/OFYAMI—have been de-identified. This includes withholding the identities of the participating school districts, and replacing students' names with unrelated numeric identifiers. As a further measure to protect confidentiality of study participants, two geographic identifiers measured at the Census Block Group (CBG) Level have been altered from their original values so that one cannot tell which CBG each participant lives in. We do so by adding a $Uniform(-k, k)$ random variable to mean CBG income and fraction of CBG minors with no health insurance, where k is set roughly at a value of 10% of the respective means ($k = 10,000$ for mean income and $k = 0.025$ for uninsured minors rate). This ensures that one cannot reverse engineer a participant's neighborhood from Census data, and clustering algorithms cannot recover groups of study participants living in the same neighborhood. On the other hand, it also largely preserves the underlying moments of the aggregate data used for analysis: perturbed versions of mean CBG income and uninsured minor rates have correlations of 0.99 with the original versions, and their correlation coefficient is 98.7% of the correlation in the raw versions of the variables. Due to this perturbation in the released CBG-level variables, the replication-file user may see small differences in the reported effects of some columns in Tables 2, 3, and 6. In principle, the information reported above can be used to deconvolve the measurement error in aggregate, while still making it impossible to recover original geographic identifiers at the individual level.

B.2. **Mathematics Pedagogical Materials.** Skill assessments and website content were comprised of professionally developed, age-appropriate math materials. We obtained copies of 46 standardized grade 5 and 6 exams used by U.S. states over the preceding decade, of which 30 were developed for 5^{th}graders and 16 were developed for 6^{th}graders: *CA Standards Test* (2009), *IL Standards Achievement Test* (2003, 2006-2011, 2013), *MN Comprehensive Assessments-Series III*, *NY State Testing Program* (2005-2010), *OH Achievement Test* (2005), *State of TX Assessments of Academic Readiness* (2011, 2013), *TX Assessment of Knowledge and Skills* (2009), and *WI Knowledge and Concepts Examinations* (2005). to develop a bank of 370 grade-5 and 302 grade-6 math questions, which we binned into five subject categories using Common Core Math Standards. We used standard Common Core subject definitions (accessible at `https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf` as of July 2025) to classify and organize pedagogical content on our website and proficiency assessments. We further categorized each problem by low, medium, and high difficulty (based on complexity and number of steps required for a solution) with help from pedagogy experts at the UChicago School Math Project. All 672 problems were pooled to expose 5^{th}- and 6^{th}-graders to the same materials.

Common Core subject definitions are grade-specific, but with considerable overlap in the concepts covered by $5^{\text{th}}$and 6^{th}graders. In Table OS.1 below we provide an overview of Common Core definitions and a harmonization of the specific subject sub-category topics that were merged to form 5-subject $5^{\text{th}}/6^{\text{th}}$grade sub-categories used in our study. The five merged subject categories are (*i*) *equations and algebraic thinking*, (*ii*) *fractions, proportions, and ratios*, (*iii*) *geometry*, (*iv*) *measurement and probability*, and (*v*) *number system*. Common

Core subject definitions for $5^{\text{th}}/6^{\text{th}}$grades differ slightly, but are overall quite similar. Our 5-subject classification is a merging of the two. Website learning tasks were organized into 25 topic-specific tasks (5 each) and 55 mixed-topic tasks with balanced portfolios of problems.

B.3. **Classroom-Based Assessments and Surveys.** Prior to randomized treatment assignment, students were given a standardized math pre-test by their teachers during regular classroom time to obtain a baseline measure of proficiency. Teachers administered a similar post-test following the experiment to gauge learning progress over the course of the study. Both assessments were designed by our research team from professionally developed, age-appropriate math materials as described in the previous section. To ensure uniformity of subject content and difficulty level, both the pre-test and post-test were populated with similar sets of 36 multiple-choice questions: 8 each from subjects (i), (iii), and (v), and 6 each from subjects (ii) and (iv). Of the 36 questions, 20 were selected from 6^{th}grade materials and the other 16 from 5^{th}grade materials, and the easy, medium, and hard categories were represented by 15, 12, and 9 questions, respectively, spread evenly across each exam. In order to discourage random answering, we instructed students that exam scores would be computed by awarding one point per correct answer, subtracting one quarter point per incorrect answer (questions had four choices), and neither adding nor subtracting points for answers left blank (contemporary SAT-exam scoring convention). Thus, students were instructed in intuitive terms to correctly solve as many problems as they could in the time provided, but to avoid randomly filling in answers on the scantron sheet. All exam scores reported and used in our empirical analyses are computed via the simple percent-correct convention.

A similarly designed post-exam was administered three weeks after the pre-exam, following the interim investment period with incentivized website activity. Following common practice for standardized exam design, there was 25% overlap in the exam questions, pre and post.

The exams were coupled with surveys to collect additional relevant information about students. Class periods were 45 minutes long; students were given 35 minutes to complete as much of the exam as they could, with the remainder of the time allocated to filling out a survey. Survey questions covered a child's attitudes and preferences (most/least favorite academic subjects and extrinsic vs. intrinsic motivation); family learning environment (# of academic helpers in the child's family/friend network and parental permissiveness for weekday video gaming and recreational internet use); and consumption/leisure options (# of video gaming systems at the child's home, fraction of peer social time under adult supervision, and enrollment in organized sports, music activities, and/or clubs). We also gathered socioeconomic indicators from the American Community Survey for each of the roughly 160 US Census block groups where students resided, our definition of a "neighborhood". For each we collected mean household income (a proxy for affluence), and the fraction of minors with no private health insurance (a proxy for deprivation of developmental resources).[1]

[1]The ACS contains many other socioeconomic indicators (e.g., mean home values) but when reported at the neighborhood level, multicollinearity problems arise due to high correlations of within-neighborhood means across different measures. We included

TABLE OS.1. Common Core Category Harmonization by Grade

GRADE 5	GRADE 6
FOCUS AREAS BY GRADE	
"In Grade 5, instructional time should focus on three critical areas: (1) developing fluency with addition and subtraction of fractions, and developing understanding of the multiplication of fractions and of division of fractions in limited cases (unit fractions divided by whole numbers and whole numbers divided by unit fractions); (2) extending division to 2-digit divisors, integrating decimal fractions into the place value system and developing understanding of operations with decimals to hundredths, and developing fluency with whole number and decimal operations; and (3) developing understanding of volume."	*"In Grade 6, instructional time should focus on four critical areas: (1) connecting ratio and rate to whole number multiplication and division and using concepts of ratio and rate to solve problems; (2) completing understanding of division of fractions and extending the notion of number to the system of rational numbers, which includes negative numbers; (3) writing, interpreting, and using expressions and equations; and (4) developing understanding of statistical thinking."*
SUB-CATEGORIES, GROUPED BY SIMILARITY	
(i) **Equations and Algebraic Thinking merged sub-category:** *"**(5OAT)** Write and interpret numerical expressions; and **(5OAT)** Analyze Patterns and Relationships."*	*(i)* **Equations and Algebraic Thinking merged sub-category:** *"**(6EE)** Apply and extend previous understandings of arithmetic to algebraic expressions; **(6EE)** Reason about and solve one-variable equations and inequalities; and **(6EE)** Represent and analyze quantitative relationships between dependent and independent variables."*
(ii) **Fractions, Proportions, and Ratios merged sub-category:** *"**(5NOF)** Use equivalent fractions as a strategy to add and subtract fractions; and **(5NOF)** Apply and extend previous understandings of multiplication and division to multiply and divide fractions."*	*(ii)* **Fractions, Proportions, and Ratios merged sub-category:** *"**(6RPR)** Understand ratio concepts and use ratio reasoning to solve problems; and **(6NS)** Apply and extend previous understandings of multiplication and division to divide fractions by fractions."*
(iii) **Geometry merged sub-category:** *"**(5GEOM)** Graph points on the coordinate plane to solve real-world and mathematical problems; and **(5GEOM)** Classify two-dimensional figures into categories based on their properties; and **(5MD)** Geometric measurement: understand concepts of volume and relate volume to multiplication and to addition."*	*(iii)* **Geometry merged sub-category:** *"**(6GEOM)** Solve real-world and mathematical problems involving area, surface area, and volume."*
(iv) **Measurement and Probability merged sub-category:** *"**(5MD)** Convert like measurement units within a given measurement system; and **(5MD)** Represent and interpret data."*	*(iv)* **Measurement and Probability merged sub-category:** *"**(6SP)** Develop understanding of statistical variability; and **(6SP)** Summarize and describe distributions."*
(v) **Number System merged sub-category:** *"**(5NOBT)** Understand the place value system; and **(5NOBT)** Perform operations with multi-digit whole numbers and with decimals to hundredths."*	*(v)* **Number System merged sub-category:** *"**(6NS)** Compute fluently with multi-digit numbers and find common factors and multiples; and **(6NS)** Apply and extend previous understandings of numbers to the system of rational numbers."*

Notes: All underlined text is the merged subject sub-categories used for our study. All italicized text is quoted from the Common Core Mathematics Standards (accessible at https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf as of July 2025). Bolded acronyms in parentheses indicate that a particular topic was taken from a given Common Core grade sub-category as follows: for grade 5, "5OAT"=*Operations and Algebraic Thinking*, "5NOBT"=*Number and Operations in Base Ten*, "5NOF"=*Number and Operations–Fractions*, "5MD"=*Measurement and Data*, and "5GEOM"=*Geometry*; for grade 6, "6RPR"=*Ratios and Proportional Relationships*, "6NS"=*The Number System*, "6EE"=*Expressions and Equations*, "6GEOM"=*Geometry*, and "6SP"=*Statistics and Probability*.

B.4. **Internet Access Issues.** Website login credentials were based on student first/last name, grade, and/or teacher name. The research team monitored a 24/7 tech-support email to resolve login issues. On the Monday following the pre-test, each participant received a personalized letter in a sealed envelope containing login credentials, website instructions, and their individual piece-rate incentive offer. A potential concern is whether students shared login credentials. This is unlikely, as work they did for someone else would offset paid work they could do for themselves. We also see a strong relationship between completed website learning tasks and improved test scores (see Tables 7 and 6), suggesting that completed website tasks reflects the user's own work.

Of the 1,676 student test subjects included in our study, 118 (7%) of them reported having no regular internet connection at home. Of these, 48 completed at least 2 learning tasks on the website, for a conditional activity rate of 40.7%. Activity rates were statistically similar for students with and without regular internet connections at home: the 95% confidence interval of this rate estimate is $(36.3\%, 45.1\%)$, which contains the activity rate for the overall sample population (44.7%). Among active students with no regular internet connection, we saw reduced rates of pageloads from desktop computers (63.5% vs 76.8%) and tablet devices (15.2% vs 18.5%), and elevated pageload rates from smartphones (21.3% vs 4.8%). The fact that the students with no regular internet connection at home still predominantly connected to our website from a personal computer is suggestive that they were able to find regular internet service elsewhere, for example in the network of public libraries serving their communities, or from the house of a family member or friend.

In order to directly test whether limited internet access played a significant role in our study, in Table OS.2 we ran two regressions of website task completion A_i on various student covariates. Specification 1 includes dummies for *no_home_internet*; school district; mean neighborhood income (a socioeconomic status proxy); self-reported regular homework time per day; *math_attitude* (a single index based on responses to preference elicitation questions on student surveys); and incentive contract dummies. In a second regression specification, we add quadratic terms for neighborhood income, homework time, and math attitude, and race/gender dummies.

Results are displayed in Table OS.2. The coefficient estimate on *no_home_internet* in both specifications is negative and of similar magnitude, but in neither is it statistically different from zero. On the other hand, other expected factors such as incentives, math attitude, and regular homework time play a significant role in predicting total website task completion. These results, and students' various outside options for connectivity (e.g., smartphones or library computers) suggest that internet access is not driving our main empirical results.

B.5. **Website Edge-Case Time Data.** There were a small number of spurious page-view times resulting when a child closed her web browser in the middle of a task without logging off. We replace these with mean work times for a student-problem-sub-category, using a

mean neighborhood income and uninsured minor rate because the two seemed most different in what they represent and had the lowest pair-wise correlation among available indicators.

Table OS.2. Determinants of Website Task Completion

(Dependent Var.: A_i)	Specification 1			Specification 2		
Regressor	**Coeff. Est.**	*(Std.Err.)*	95% Conf. Int.	**Coeff. Est.**	*(Std.Err.)*	95% Conf. Int.
no_home_internet	-2.47	*(1.939)*	[-6.27,1.33]	-2.39	*(1.949)*	[-6.21,1.43]
*District*2	-9.28***	*(1.616)*	[-12.44,-6.11]	-7.15***	*(2.046)*	[-11.16,-3.14]
*District*3	-14.50***	*(2.801)*	[-19.99,-9.01]	-11.03***	*(3.526)*	[-17.94,-4.12]
nbhd_income	-2.2×10^{-5}	(2.8×10^{-5})	$[-7.7,3.2]\times10^{-5}$	1.0×10^{-5}	(9.5×10^{-5})	$[-1.8,2.0]\times10^{-4}$
$nbhd\_income^2$	—			-1.7×10^{-10}	(4.2×10^{-10})	$[-9.9,6.6]\times10^{-10}$
$(hmwk\_time/day)$	-1.71	*(1.168)*	[-4.00,0.58]	1.97	*(2.842)*	[-3.60,7.54]
$(hmwk\_time/day)^2$	—			-2.09**	*(0.934)*	[-3.93,-0.26]
math_attitude	2.52***	*(0.391)*	[1.75,3.29]	1.36***	*(0.489)*	[0.41,2.32]
$math\_attitude^2$	—			0.38***	*(0.096)*	[0.19,0.57]
*Contract*2	2.48**	*(1.226)*	[0.08,4.89]	2.43**	*(1.220)*	[0.03,4.82]
*Contract*3	4.70***	*(1.225)*	[2.30,7.10]	4.67***	*(1.217)*	[2.27,7.04]
Constant	12.23***	*(4.062)*	[4.27,20.19]	8.41	*(6.244)*	[-3.83,20.65]
Gender/Race Dummies	NO			YES		
R^2	0.124			0.141		

Notes: Statistical significance at the 10%, 5%, and 1% levels are denoted by "*," "**," "***," respectively.

procedure proposed by (Cotton et al., 2022). Other "edge cases" (atypical user behavior) included when users re-attempted a module after having passed it, or when they attempted some module but never completed it. Time accrued in either scenario is labeled "unpaid time," and we experimented with 3 methods for handling it. *Method 1* aggregates all of child j's unpaid time and adds an equal portion of it to each work-time log τ_{ja} for completed modules. In *Method 2* we apply unpaid time serially to success times in progress. E.g., suppose j sequentially passes module (i) in 15 minutes, then spends 2 minutes re-attempting module (i), followed by 6 minutes attempting module (v) unsuccessfully, followed by 7 minutes attempting module (iv) with a success. Assuming j never completes module (v), her first work time log is $\tau_{j1}=15$ (with $a=1$ representing her first success on module (i)), and her second work time is $\tau_{j2}=2+6+7=15$ (with $a=2$ representing her second success on module (iv)). Finally, *Method 3* simply drops all unpaid time. Our empirical implementation used *Method 2*, which implies the interpretation that t^*_{ja} is the maximal time j can rationally spend in pursuit of an a^{th} success, after $(a-1)$ previous successes. The correlation of time logs across methods *1* and *2* is 0.998, and across methods *2* and *3* is 0.993 (sample size: 16,380 child-completion pairs). The high correlations are due to the fact that edge cases make up a small fraction of total logged time on the website.

Appendix C. Structural Estimator Technical Details

C.1. **Stage 1: Differencing Estimator and Asymptotic Properties.** Here we cover technical details on our differencing estimator $\hat{\boldsymbol{B}}_p$ mentioned in Section 4.1. We implement a standard approach for unbalanced panel data as developed in Wooldridge (2002) (Chapter 17). The cross-section sampling index is $i=1,\dots,N$, and the chronological task-completion index a ranges between values of 2 and 80 for active students. For simplicity, in this section we refer to a as the "time" index, and we define $\boldsymbol{x}_{ia}=(\mathbb{1}(a=1),\log(a))$ as the a-specific regressors from equation (5). Recall that we have an unbalanced panel issue: letting $\boldsymbol{\mathcal{S}}_i \equiv (\mathcal{S}_{i1}, \ldots, \mathcal{S}_{i80})'$ denote the 80×1 vector of selection indicators, where $\mathcal{S}_{ia} = 1$ if $2 \leq a \leq A_i$, so that $(\boldsymbol{x}_{ia}, \tau_{ia})$ is observed, and zero otherwise. Let $\boldsymbol{\mathcal{B}}_p \equiv [\log(\mathcal{T}_1), \varphi]'$ denote the vector of slope coefficients from $\boldsymbol{B}_p$. With that, $\hat{\boldsymbol{\mathcal{B}}}_p$ is obtained as an OLS regression on the transformed data:

$$\begin{aligned}\hat{\boldsymbol{\mathcal{B}}}_p &= \left(N^{-1}\sum_{i=1}^{N}\sum_{a=a_0}^{80}\mathcal{S}_{ia}\ddot{\boldsymbol{x}}'_{ia}\ddot{\boldsymbol{x}}_{ia}\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\sum_{a=a_0}^{80}\mathcal{S}_{ia}\ddot{\boldsymbol{x}}'_{ia}\ddot{\log}(\tau_{ia})\right)\\ &= \boldsymbol{\mathcal{B}}_p + \left(N^{-1}\sum_{i=1}^{N}\sum_{a=a_0}^{80}\mathcal{S}_{ia}\ddot{\boldsymbol{x}}'_{ia}\ddot{\boldsymbol{x}}_{ia}\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\sum_{a=a_0}^{80}\mathcal{S}_{ia}\ddot{\boldsymbol{x}}'_{ia}\log(u_{ia})\right),\end{aligned} \tag{12}$$

where $\ddot{\boldsymbol{x}}_{ia} \equiv \boldsymbol{x}_{ia} - A_i^{-1}\sum_{a=a_0}^{80}\mathcal{S}_{ia}\boldsymbol{x}_{ia}$, $\ddot{\log}(\tau_{ia}) \equiv \log(\tau_{ia}) - A_i^{-1}\sum_{a=a_0}^{80}\mathcal{S}_{ia}\log(\tau_{ia})$, and $a_0 \equiv 1$ if using the time de-meaning method.[2]

Assumption 4 implies that $\mathrm{E}[\log(U_{ia})|\boldsymbol{x}_i, \boldsymbol{\mathcal{S}}_i, \log(\theta_{pi})] = 0$ since U_{ia} is independent of a and θ_{mi}, and since $\log(U_{ia})$ is mean independent of $\log(\theta_{pi})$. Thus, from the law of iterated expectations it follows that $\mathrm{E}[\mathcal{S}_{ia}\ddot{\boldsymbol{x}}'_{ia}\log(u_{ia})] = \boldsymbol{0}$ (because $\mathcal{S}_{ia}\ddot{\boldsymbol{x}}_{ia}$ is a function of $(\boldsymbol{x}_i, \mathcal{S}_i)$). Recall that the selection indicators $\boldsymbol{\mathcal{S}}_i$ encapsulate student decisions of whether or not to complete incentivized learning tasks on the website. Assumption 4 essentially states that these decisions are based on stable unobserved types $(\theta_{pi}, \theta_{mi})$, rather than on the average value of the transitory shocks $\log(U_{ia})$. Verbeek and Nijman (1992) propose a test for selectivity bias in panel-data settings as follows:

- ***step (i)*** add a lead selection indicator, $\mathcal{S}_{i,a+1}$ to the vector $\boldsymbol{x}_{ia}$;
- ***step (ii)*** estimate the model by fixed effects (dropping all i such that $A_i = 2$, and dropping the last period for all i) as outlined above; and
- ***step (iii)*** run a robust t-test for the null hypothesis that the coefficient on the lead selection indicator is zero.

Under this null hypothesis, $log(U_{ia})$ is uncorrelated with $\mathcal{S}_{ir}$ for all r. We execute the Verbeek-Nijman test and get a t-statistic of -0.374 (df=15,984) and a p-value of 0.709, meaning we fail to find evidence contradicting the conditions of Assumption 4.

Under the conditions of Assumption 4, if we define $\boldsymbol{z}_i \equiv \sum_{a=1}^{80}\mathcal{S}_{ia}\ddot{\boldsymbol{x}}'_{ia}\log(u_{ia})$ then $E[\boldsymbol{Z}] = \boldsymbol{0}$, and consistency follows from the weak law of large numbers (as $N \to \infty$) since the last term of the second line of equation (12) is just a sample mean of the random variable $\boldsymbol{Z}$, and from the continuous mapping theorem since $\mathcal{S}_{ia}\ddot{\boldsymbol{x}}'_{ia}\ddot{\boldsymbol{x}}_{ia}$ is (by construction) non-singular within the unbalanced panel (i.e., $\forall i$ such that $A_i \geq 2$). Asymptotic normality follows from similar logic combined with the Lindeberg-Levy central limit theorem.

Moreover, given consistency of $\hat{\boldsymbol{\mathcal{B}}}_p$, by the continuous mapping theorem we can define a consistent estimator of the log-shocks as $\widehat{\log(u_{ia})} \equiv \ddot{\log}(\tau_{ia}) - \ddot{\boldsymbol{x}}_{ia}\hat{\boldsymbol{\mathcal{B}}}_p - \widehat{\log(\mathcal{T}_0)}$ for each (i,a) pair such that $\mathcal{S}_{ia} = 1$, where $\widehat{\log(\mathcal{T}_0)} \equiv \frac{\sum_{i=1}^{N}\sum_{a=1}^{80}\mathcal{S}_{ia}[\ddot{\log}(\tau_{ia}) - \ddot{\boldsymbol{x}}_{ia}\hat{\boldsymbol{\mathcal{B}}}_p]}{\sum_{i=1}^{N}\sum_{a=1}^{80}\mathcal{S}_{ia}}$ and $\hat{\boldsymbol{B}}_p = [\widehat{\log(\mathcal{T}_0)}, \hat{\boldsymbol{\mathcal{B}}}'_p]'$. Finally,

[2] Alternatively, one could use the time differencing method by transforming the regressors as $\ddot{\boldsymbol{x}}_{ia} \equiv \mathcal{S}_{ia}\boldsymbol{x}_{ia} - \mathcal{S}_{i,a-1}\boldsymbol{x}_{i,a-1}$ and the outcomes as $\ddot{\log}(\tau_{ia}) \equiv \mathcal{S}_{ia}\log(\tau_{ia}) - \mathcal{S}_{i,a-1}\log(\tau_{i,a-1})$, for $a \geq a_0 \equiv 2$.

by equation (5) and the continuous mapping theorem, we also have that the following is a $\sqrt{A_i}$-consistent estimator of log-productivity types for all i such that $A_i \geq 2$:[3]

$$\widehat{\log(\theta_{pi})} \equiv \frac{\sum_{a=1}^{A_i} \left\{ \log(\tau_{ia}) - \widehat{\log(\mathcal{T}_0)} - \widehat{\log(\mathcal{T}_1)} \mathbb{1}(a=1) - \hat{\varphi} \log(a) - \widehat{\log(u_{ia})} \right\}}{A_i}. \tag{13}$$

C.2. **Stage 2: Method of Simulated Moments (MSM) Estimator Setup.** Given our cubic B-Spline cost function, $\hat{c}(t;\boldsymbol{\gamma}_c) = \sum_{k=1}^{K_c+3} \gamma_{ck} \mathcal{B}_{ck}(t)$, with knot vector $\boldsymbol{\kappa}_c = \{\kappa_{c1}, \kappa_{c2}, \ldots, \kappa_{c,K_c+1}\}$ and $K_c = 7$,[4] our estimator is built on functional representations of counterfactual quantile comparisons embedded within the identification proof in Section 4.2. For fixed $\boldsymbol{\gamma}_c$ and for each contract j, we can compute empirical and model-based labor-supply as follows:

(*i*) Simulate sequences of work-time shocks $\left\{ \{\tilde{u}_{iaq}\}_{a=1}^{\overline{A}_i} \right\}_{q=1}^{Q}$ from $\hat{F}_u(u|\hat{\theta}_{pi})$, given $\hat{\theta}_{pi}$, $\hat{\mathcal{T}}_0$, $\hat{\mathcal{T}}_1$, and $\hat{\varphi}$, where sequence length is chosen as $\overline{A}_i = \max\{40, \min\{160, 2A_i\}\}$.

(*ii*) For all i assigned to contract j, we find $\theta_{mi}(\boldsymbol{\gamma}_c)$ such that the sample mean of simulated optimal stop times $\{\tilde{T}_{ijq} = L_T(\hat{\theta}_{pi}, \theta_{mi}(\boldsymbol{\gamma}_c), \pi_{1j}, \{\tilde{u}_{iaq}\}_{a=1}^{\overline{A}}; \boldsymbol{\gamma}_c)\}_{q=1}^{Q}$ equals T_i.[5]

(*iii*) For each i assigned to contract $j = 1, 2, 3$, compute counterfactual labor supply, $\{\tilde{A}_{ij'q} = L_A(\hat{\theta}_{pi}, \theta_{mi}(\boldsymbol{\gamma}_c), \pi_{1j'}, \{\tilde{u}_{iaq}\}_{a=1}^{\overline{A}}; \boldsymbol{\gamma}_c)\}_{q=1}^{Q}$ under each alternate contract $j' \neq j$.[6]

(*iv*) For $j' = 1, 2, 3$, compute model-generated labor-supply CDFs $\tilde{G}_A(a|\pi_{1j'}; \boldsymbol{\gamma}_c) = \sum_{i=1}^{N} \sum_{q=1}^{Q} \frac{\mathbb{1}\left[\tilde{A}_{ij'q} \leq a \ \& \ \pi_1^i \neq \pi_{1j'}\right]}{\sum_{i=1}^{N} \mathbb{1}\left[\pi_1^i \neq \pi_{1j'}\right] \times Q}$ under assignment to contract j'.

(*v*) For $j = 1, 2, 3$, compute empirical labor-supply CDFs, $\hat{G}_A(a|\pi_{1j}) = \sum_{i=1}^{N} \frac{\mathbb{1}\left[A_i \leq a \ \& \ \pi_1^i = \pi_{1j}\right]}{\sum_{i=1}^{N} \mathbb{1}\left[\pi_1^i = \pi_{1j}\right]}$.

Thus, i's observed choices (T_i, A_i) contribute to the empirical CDF of her assigned group, $\hat{G}_A(a|\pi_{1j})$, and they also contribute to the counterfactual CDFs $\tilde{G}_A(a|\pi_{1j'}; \boldsymbol{\gamma}_c)$, $j \neq j'$, through $\theta_{mi}(\boldsymbol{\gamma}_c)$. To form the MSM objective, all CDF values are linearly interpolated on a grid $\{a_1, a_2, \ldots, a_L\} \subset [2, 80]$, and the cost parameter estimator $\hat{\boldsymbol{\gamma}}_c$ minimizes the distance between the empirical CDFs and their model-generated counterparts:

$$\begin{aligned} \hat{\boldsymbol{\gamma}}_c &= \operatorname{argmin} \left\{ \sum_{l=1}^{L} \sum_{j=1}^{3} \left(\hat{G}_A(a_l|\pi_{1j}) - \tilde{G}_A(a_l|\pi_{1j}; \boldsymbol{\gamma}_c) \right)^2 \right\} \\ &s.t.\ \gamma_{c1} = 0;\ \gamma_{c2} = (\kappa_{c2} - \kappa_{c1})/3;\ \gamma_{ck} - \gamma_{c,k-1} > 0,\ k = 2, \ldots, K_c + 3; \\ &\frac{\gamma_{c,k+1} - \gamma_{ck}}{\kappa_{c,k+1} - \kappa_{ck}} - \frac{\gamma_{ck} - \gamma_{c,k-1}}{\kappa_{ck} - \kappa_{c,k-1}} > 0\ k = 2, \ldots K_c + 2. \end{aligned} \tag{14}$$

[3]The consistency concept is well-posed, provided that the sampling process could be hypothetically altered to increase the relevant sample size. $\widehat{\boldsymbol{B}}_p$ is consistent in the cross-section size N, as a research team with unbounded resources could hypothetically recruit more test subjects. Similarly, given unbounded research budget one could extend the time horizon (e.g., 6 weeks or more) or increase monetary incentives (e.g., $\pi_1 \in \{\$3, \$4, \$5, \ldots\}$), either of which would increase panel length A_i.

[4]We chose $K_c = 7$, or 8 knots (see appendix C.3). After imposing Assumption 2, $\boldsymbol{\gamma}_c$ contains 8 remaining free parameters.

[5]In a slight shift in notation, here we use T_i to denote i's total work time through all *completed* learning tasks. While this leaves a small amount of empirical information on the table, it drastically reduces the number of required continuation-value evaluations. When solving for optimal stopping choices, we employ a new method recently proposed by Hamilton, Hickman, and Mohie (2025) for tractable computation of non-stationary dynamic programming problems with history dependence.

[6]It is computationally advantageous to simulate counterfactual labor supply in A units, as this eliminates many continuation-value evaluations required to compute T_i, including work times for terminal units of work left incomplete.

Constraints enforce boundary value ($c(0)=0$), boundary derivative ($c'(0)=1$), monotonicity ($c'(t)>0$), and convexity ($c''(t)>0$). Imposing shape conditions on B-Splines requires only linear constraints on the parameters, which improves numerical solver properties.

To aid estimator computation, we augment the MSM objective (14) with two additional "guardrail" terms: letting $\hat{\overline{G}}^{90}_a(a_l|\pi_{1j})$ and $\hat{\underline{G}}^{90}_a(a_l|\pi_{1j})$ denote the interpolated point-wise 90% confidence bounds of the empirical CDFs, we add the following for each l in the outer sum

$$\omega_0 \times \left(\hat{\overline{G}}^{90}_a(a_l|\pi_{1j}) - \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right)^2 \times \mathbb{1}\left[\hat{\overline{G}}^{90}_a(a_l|\pi_{1j}) < \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right] \quad \text{and}$$

$$\omega_0 \times \left(\hat{\underline{G}}^{90}_a(a_l|\pi_{1j}) - \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right)^2 \times \mathbb{1}\left[\hat{\underline{G}}^{90}_a(a_l|\pi_{1j}) > \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right],$$

where ω_0 is a penalty parameter. These guardrail conditions help the solver avoid iterating through regions of $\boldsymbol{\gamma}_c$-space with poor model fit. They impose an extra quadratic penalty only if the model-generated CDFs differ from their empirical analogs by more than the 90% confidence bounds. They were included to reduce the number of iterations required for solver convergence, which is an important concern when nested fixed-point solution methods are unavoidable. Figure 6 shows that once the solver converged, the model-projected CDFs lay inside the 90% confidence bounds, so guardrail conditions are inactive at the solution. Thus, they can be ignored for the discussion on asymptotics.

C.2.1. *Stage-2: Dealing With Mass Points at the Sample Extrema.* We have a small mass of students who achieve full output $A_i=80$ on the website, as can be seen in Figure 1. This means that their study-time productivity trait, θ_{pi}, is known, but without extra structure their motivation trait, θ_{mi}, can only be bounded from above. This is because it is impossible to know whether a given individual would have optimally chosen *exactly* $A_i=80$, or $A_i>80$.[7] We deal with this problem by estimating a constrained quantile function using a low-dimensional B-spline to extrapolate into the missing upper tails of the empirical CDFs of A. The extrapolating B-spline quantile functions overlapped their empirical counterparts to the 85th percentile. We assumed that no student would choose to more than double the available workload on the website, so tails were bounded from above by $A=160$. We chose a low-dimensional B-spline with 3 knots so that all parameters for the extrapolating quantile functions could be informed by the available data. One advantage of this approach is that we can pre-estimate the extrapolated upper tails of the work volume distributions, without adding to the computational complexity of the main simulated GMM estimator.

We discretized the extrapolated tails (for computational tractability) by selecting no more than 5 uniform steps (in quantile rank space), and also requiring each step (except possibly the last one) to represent at least 5 observations of $A_i = 80$. The resulting frequency tables included 3 steps under contract 1 (with the smallest upper mass point), and 5 steps each for contracts 2 and 3. Figure OS.3 in the online supplement plots the extrapolated upper tails against the empirical CDFs of A. After discretizing the upper tail, for each

[7] Note, however that this bound is much tighter than the bounds on Marginal/Inactive student motivation types.

individual with full output this renders up to 5 possibilities for optimal stopping points $\{\widehat{A}_{i1},\ldots,\widehat{A}_{i5}\}$, all being at or above 80. For each $(\theta_{pi},\widehat{A}_{im})$ pair, $m=1,\ldots,5$, we back out a motivation trait $\theta_{mi}(\widehat{A}_{im})$ to match $\widehat{A}_{im}$ as the optimal stopping point, and we run counterfactual simulations for each $(\theta_{pi},\theta_{mi}(\widehat{A}_{im}))$ pair. However, we give each of these $(1/5)^{\text{th}}$ weight when incorporating them into the model-generated CDFs $\widetilde{G}_a$.

The other minor challenge relates to non-active students who completed fewer than 2 learning tasks $A_i \le 1$ (i.e., upper tail of the (θ_p,θ_m) distribution). This is not a threat to structural identification, which relies only on exogenous incentive variation across comparable samples of students. Recall from Table 1 that masses of active students, $M_j^{act}\equiv N_j^{act}/N_j$ in Contract Groups $j=1,2,3$ were ordered as follows: $M_1^{act}<M_2^{act}<M_3^{act}$. Thus, within the first two groups there were fractions $\frac{M_3^{act}-M_1^{act}}{M_3^{act}}$ and $\frac{M_3^{act}-M_2^{act}}{M_3^{act}}$ of students who would have entered active status $(A_i\ge 2)$ under contract-3 incentive π_{13}. For contract groups 1 and 2 we compute simulated counterfactual choices for marginal students (i.e., $A_i=1$), and we assign weight $\omega_j\equiv\frac{M_3^{act}-M_1^{act}}{M_3^{act}}\cdot\frac{N_j^{act}}{N_j^{mrg}}$, $j=1,2$, to their simulated counterfactual choices when computing the model-generated CDFs. This coping strategy ensures that the lower tails of empirical CDFs, $\hat{G}_A(A|\pi_{1j})$, and their model-generated counterfactual analogs, $\tilde{G}_A(a|\pi_{1j})$, $a\ge 2$, are based on underlying sets of students with comparable unobserved types.

C.2.2. *Stage-2 Asymptotic Properties.* Regarding asymptotic theory, recall that the only free parameters being pinned down in the Stage-2 MSM objective are $\boldsymbol{\gamma}_c$, with motivation fixed effects $\theta_{mi}(\boldsymbol{\gamma}_c)$ being recovered from the solution of equation (2), given $\boldsymbol{\Omega}_{ij}$ and observed T_i. Treating the cubic B-spline specification $\hat{c}(t;\boldsymbol{\gamma}_c)=\sum_{k=1}^{K_c+3}\gamma_{ck}\mathcal{B}_{ck}(t)$, $K_c=7$, as a fixed parametric form, standard asymptotic theory for MSM estimators applies (see Davidson and MacKinnon (2021), Section 9.6). The estimator $\widehat{\boldsymbol{\gamma}}_c$ defined in (14) is $\sqrt{N}$-consistent and asymptotically normal under the usual regularity conditions: (i) the model is identified (see Section 4.2), and (ii) the model is correctly specified, which appears suggestively plausible in Figure 6 of Appendix A. Finally, consistency of $\hat{\theta}_{mi}$ follows by the continuous mapping theorem and by consistency of all the objects it is a function of, but since it depends on the $\sqrt{A_i}$-consistent $\hat{\theta}_{pi}$ for each active student i, it is similarly consistent in panel length A_i.

C.3. **B-Spline Function Specification.** While B-splines may produce a semi-nonparametric sieve—if one allows K_c to grow with N—asymptotic properties of such an approach are an open question due to the non-nested nature of successive B-spline models. Thus, we view our 10-parameter B-spline cost function as a flexible but fixed parametric form. We do not formally optimize B-spline model selection (i.e., choice of knot vectors) via cross validation for various reasons. This would create formidable applied mathematical challenges stemming from the computationally expensive nested-fixed-point approach to solving the MSM objective in our Stage-2 structural estimator. This makes it burdensome to repeatedly re-estimate, a challenge that would be further compounded by the need to optimize a high-dimensional tuning parameter, including number of knots and knot locations. The cost is that we potentially give up on some efficiency, but the benefit is that we are relieved of a prohibitive computational burden. To do cross-validation would also require sample splitting, which would mitigate potential efficiency gains by reducing the size of the estimation set.

Instead, we chose knot vectors by a simple rule of thumb to place knots at the quantiles of the data. This convention ensures that the collective influence of the data is spread evenly across the B-spline parameters, and it reduces the tuning parameter choice to a single dimension of how many knots. For the shock distribution CDFs F_u, we placed knots at the terciles (i.e., 4 knots each) of the log-u_{ia} samples corresponding to each interval of θ_p. This resulted in excellent model fit due to the remarkable flexibility of B-splines, but is also due to the fact that the empirical CDF shapes were fairly simple (see Figure OS.2).

For the cost function $c(t; \boldsymbol{\gamma}_c)$, which had to be able to replicate the highly skewed distributions depicted in Figure 1, we placed knots at the septiles (i.e., double the amount of knots at 8, rather than 4). The main goal was to ensure that model choice did not unduly restrict cost curvature, so that the data could directly determine the shape of $c(t; \boldsymbol{\gamma}_c)$ to the greatest extent possible. Two indicators suggest that this goal was achieved. First, model fit depicted in Figure 6 is good, with model-generated CDFs all within the 90% confidence bounds of their empirical analogs. Second, the estimated cost function in Figure 2 displays significantly more curvature than linear costs (positive second derivative constraints were were imposed), but still much less curvature at each point than what a cubic B-spline with 8 knots is capable of accommodating. In that sense, our cost function estimate $c(t; \hat{\boldsymbol{\gamma}}_c)$ is within the interior of the space of functions spanned by our B-spline cost specification.

C.4. **Stages 1& 2 Standard Errors.** We bootstrap standard errors for estimates of Stages 1 and 2. Our block-bootstrap procedure mimics our randomized sampling procedure (discussed in Section 3.2.3) which balanced on race, gender, school district, grade level, and pre-test score. We begin by arranging all test subjects into race-gender-district-grade bins.[8] Suppose that there are K such bins in total, and that within contract $j = 1, 2, 3$ the bins each have $N_{1j},\ N_{2j}, \ldots, N_{Kj}$ subjects in them, respectively. Then, in order to construct a bootstrap sample, for each bin $k = 1, \ldots, K$ we do the following:

(1) Randomly draw a test subject (with replacement), call her "$subject_1$," and record which contract j she was assigned.
(2) Select subjects from the other two contracts j' and j'' in that same race-gender-district-grade bin (with replacement) whose pre-test scores are closest to $subject_1$'s pre-test score. Break ties randomly if multiple subjects fit that description within contract groups j' and/or j''. Call these two selected individuals "$subject_2$" and "$subject_3$,"respectively.
(3) Add the triple $(subject_1, subject_2, subject_3)$ to the bootstrap sample.
(4) Repeat steps (1)–(3) above, until full bootstrap samples of size N_{k1}, N_{k2}, and N_{k3} have been constructed for bin k under contracts 1, 2, and 3, respectively.

[8] Due to a sparsity of Black and Hispanic students in District 1 and a sparsity of White and Asian students in District 3, we only arrange students into gender-district-grade bins in those two districts. District 2 subjects, who exhibit a more diverse racial mix, are fully partitioned into race-gender-district-grade bins.

(5) Repeat steps (1)–(4) above for each race-gender-district-grade bin, $k = 1, \ldots, K$.

The final remaining question is how many bootstrap samples on which to generate and re-estimate the model. The main consideration here is a trade-off between simulation error and computational cost. Estimates of the student time allocation model generally took between 20 and 90 minutes each, including an adaptive multiple re-start algorithm to ensure quality of each final solution, so we chose 400 total bootstrap samples.

For standard errors on student fixed effects, we first bootstrap all common parameters. Then, we combine the bootstrapped parameter samples, $\left\{\mathcal{T}_0^{(s)}, \mathcal{T}_1^{(s)}, \varphi^{(s)}, \boldsymbol{\gamma}_c^{(s)}\right\}_{s=1}^{400}$, etc., with an individual's observables, $\left\{\{\tau_{ia}\}_{a=1}^{A_i}, T_i, A_i, \boldsymbol{X}_{pi}, \boldsymbol{X}_{mi}\right\}$, to compute bootstrapped fixed effect estimates $\left\{\theta_{pi}^{(s)}, \theta_{mi}^{(s)}\right\}_{s=1}^{400}$. These within-student bootstrap samples of fixed effects are then used to compute standard errors, inverse variance weights, and EB shrinkage forecasts.

C.5. **Stage 3: Tobit Estimator Technical Details.** We calculate simulated selection probabilities as follows. We simulate a 1,000×2 matrix $\boldsymbol{Z}$ of independent standard normal shocks. For each marginal/inactive student i, and for each guess k of the parameters $(\boldsymbol{\beta}_{pk}, \boldsymbol{\beta}_{mk}, \boldsymbol{\Sigma}_k)$, we transform the shocks by $(\tilde{\boldsymbol{\eta}}_{pki}, \tilde{\boldsymbol{\eta}}_{mki}) = \boldsymbol{V}_i'\boldsymbol{Z}$, where $\boldsymbol{V}_i$ is the (upper) Cholesky decomposition of $\boldsymbol{\Sigma}_{ki}$, to impose the BVN correlation structure. We add the mean-shifters $\boldsymbol{X}_{ji}\boldsymbol{\beta}_{jk}$, $j = p, m$, to get a 1,000×2 sample of simulated types $(\tilde{\boldsymbol{\theta}}_{pki}, \tilde{\boldsymbol{\theta}}_{mki})$, using equation (7), for selection probabilities in (9). One might use a sample mean of indicators, $\Pr\left[\theta_m > \underline{\Theta}_m(\theta_p; b_i, \pi_{1i}, \widehat{\gamma}_c)\middle| \boldsymbol{X}_{pi}, \boldsymbol{X}_{mi}; \boldsymbol{\beta}_p, \boldsymbol{\beta}_m, \boldsymbol{\Sigma}_i\right] = \sum_{s=1}^{1{,}000} \mathbb{1}\left(\tilde{\theta}_{mkis} \geq \underline{\Theta}_m(\tilde{\theta}_{pkis}; \pi_0^i, \pi_1^i, \widehat{\gamma}_c)\right)/1{,}000$, but this would introduce discontinuities into the Tobit objective function. Thus, we instead compute smoothed selection probabilities as $\sum_{s=1}^{1{,}000} K\left(\frac{\tilde{\theta}_{mkis} - \underline{\Theta}_m(\tilde{\theta}_{pkis}; \pi_0^i, \pi_1^i, \widehat{\gamma}_c)}{h}\right)/1{,}000$, where $K(\cdot)$ is a triweight kernel CDF (with density levels and slopes equaling zero at the support bounds), and bandwidth h is chosen by Silverman's rule of thumb applied to the sample of identified $\hat{\theta}_m$ types for active students: $h = 1.06(2.978)\widehat{StDev}(\theta_m | A \geq 2)\left(\sum_{i=1}^{N} \mathbb{1}(A_i \geq 2)\right)^{-1/5}$.

Tobit ML and simulation-based M-estimators are known to exhibit problems of local optima, so we also employ a numerical solution strategy that combines various quasi-Newton and derivative-free solvers, along with an extensive array of multiple re-starts, to ensure convergence to a global optimum. Here we outline our approach to numerically solving the Tobit ML estimator from Section 5. In order to ensure convergence to a global optimum, we adopted the following approach to solving for the arg max of the Tobit objective function (9), given an initial guess of the parameters $(\boldsymbol{\beta}_{p0}, \boldsymbol{\beta}_{m0}, \boldsymbol{\Sigma}_0)$:

(1) Let a candidate solution be denoted by $(\hat{\boldsymbol{\beta}}_p^*, \hat{\boldsymbol{\beta}}_m^*, \hat{\boldsymbol{\Sigma}}^*, \mathcal{L}^*)$, where $\mathcal{L}^*$ is its corresponding log-likelihood function value. Iteratively obtain this candidate solution as follows: solve for $(\hat{\boldsymbol{\beta}}_{p,k+1}, \hat{\boldsymbol{\beta}}_{m,k+1}, \hat{\boldsymbol{\Sigma}}_{k+1}, \mathcal{L}_{k+1})$ 9 times, using $(\hat{\boldsymbol{\beta}}_{pk}, \hat{\boldsymbol{\beta}}_{mk}, \hat{\boldsymbol{\Sigma}}_k, \mathcal{L}_k)$ as an initial guess, and various MATLAB (v2022b) solvers in the following order:
   (a) $k = 0$: (quasi-Newton) interior-point algorithm;
   (b) $k = 1$: (quasi-Newton) sequential quadratic programming legacy (SQP-L) algorithm;

(c) $k=2$: (quasi-Newton) sequential quadratic programming (SQP) algorithm;
(d) $k=3$: (quasi-Newton) active-set algorithm;
(e) $k=4$: (derivative-free) Non-Uniform Pattern Search algorithm with mesh-adaptive direct search option (NUPS-MADS);
(f) $k=5$: (derivative-free) Uniform Pattern Search algorithm;
(g) $k=6$: (derivative-free) Non-Uniform Pattern Search, default algorithm;
(h) $k=7$: (derivative-free) Non-Uniform Pattern Search algorithm with generalized pattern search option;
(i) $k=8$: (quasi-Newton) interior-point algorithm;

specify $(\hat{\boldsymbol{\beta}}^*_p, \hat{\boldsymbol{\beta}}^*_m, \hat{\boldsymbol{\Sigma}}^*, \mathcal{L}^*) = (\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9, \mathcal{L}_9)$.

(2) Generate multiple re-start values to re-solve from as follows:

(a) Define seven variable groups (*I*) neighborhood SES controls; (*II*) race, gender, age; (*III*) peer/adult helpers, math attitudes, extrinsic/intrinsic scores; (*IV*) #gaming systems, weekday gaming permission, weekday recreational internet permission, fraction leisure time under adult supervision, sports/music/clubs enrollment; (*V*) study time, screen time, home connectivity controls; (*VI*) parent survey controls; (*VII*) school district dummies.

(i) For each variable group G, find three *holdout restart points* $(\boldsymbol{\beta}^{HGl}_p, \boldsymbol{\beta}^{HGl}_m, \boldsymbol{\Sigma}^{HGl})$, $l=1,2,3$, by the following:

(A) $l=1$: optimize (9) using $(\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9)$ as a start point, but constraining group G coefficients in the θ_p equation only to be zero;

(B) $l=2$: optimize (9) using $(\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9)$ as a start point, but constraining group G coefficients in the θ_m equation only to be zero;

(C) $l=3$: optimize (9) using $(\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9)$ as a start point, but constraining group G coefficients in both equations to be zero.

(D) For each case above, iteratively solve the constrained Tobit objective twice using the interior-point and SQP-L algorithms in that order.

(ii) For each variable group G, find 12 *scaled restart points* $(\boldsymbol{\beta}^{SGl}_p, \boldsymbol{\beta}^{SGl}_m, \boldsymbol{\Sigma}^{SGl})$, $l=1,\ldots,6$, by the following:

(A) $l=1,2,3$: using $(\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9)$ as a start point, scale group G coefficients in the θ_p equation only by a factor of (1/2); then do the same for group G coefficients in the θ_p equation only; then do the same for group G coefficients in both equations.

(B) $l=4,5,6$: using $(\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9)$ as a start point, scale group G coefficients in the θ_p equation only by a factor of 2; then do the same for group G coefficients in the θ_p equation only; then do the same for group G coefficients in both equations.

(C) $l=7,8,9$: using $(\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9)$ as a start point, scale group G coefficients in the θ_p equation only by a factor of (1/4); then do the same for group G coefficients in the θ_p equation only; then do the same for group G coefficients in both equations.

(D) $l = 10, 11, 12$: using $(\hat{\boldsymbol{\beta}}_{p9}, \hat{\boldsymbol{\beta}}_{m9}, \hat{\boldsymbol{\Sigma}}_9)$ as a start point, scale group G coefficients in the θ_p equation only by a factor of 4; then do the same for group G coefficients in the θ_p equation only; then do the same for group G coefficients in both equations.

(b) For each generated re-start point—that is, for $(\boldsymbol{\beta}_p^{rGl}, \boldsymbol{\beta}_m^{rGl}, \boldsymbol{\Sigma}^{rGl})$, $r=H,S$ and for each corresponding l— obtain candidate solution $(\boldsymbol{\beta}_p^{rGl*}, \boldsymbol{\beta}_m^{rGl*}, \boldsymbol{\Sigma}^{rGl*}, \mathcal{L}^{rGl*})$ by iteratively re-solving the unconstrained Tobit objective function (9) four times using $(\boldsymbol{\beta}_p^{rGl}, \boldsymbol{\beta}_m^{rGl}, \boldsymbol{\Sigma}^{rGl})$ as a start point and the following order of solvers

(i) $k = 0$: (quasi-Newton) interior-point algorithm;

(ii) $k = 1$: (quasi-Newton) SQP algorithm;

(iii) $k = 2$: (derivative-free) NUPS-MADS algorithm;

(iv) $k = 3$: (derivative-free) Uniform Pattern Search algorithm.

(c) Define $(\boldsymbol{\beta}_p^{**}, \boldsymbol{\beta}_m^{**}, \boldsymbol{\Sigma}^{**}, \mathcal{L}^{**}) = \left\{ (\boldsymbol{\beta}_p^{rGl*}, \boldsymbol{\beta}_m^{rGl*}, \boldsymbol{\Sigma}^{rGl*}, \mathcal{L}^{rGl*}) : \mathcal{L}^{rGl*} = \min_{r'=H,S;G'=I,\dots,VII;l'} \{\mathcal{L}^{r'G'l'*}\} \right\}$.

(3) If $\mathcal{L}^{**} > \mathcal{L}^* + tol$—that is, if the re-start solutions collectively achieve a non-trivial improvement over the original candidate solution, re-define $(\boldsymbol{\beta}_{p0}, \boldsymbol{\beta}_{m0}, \boldsymbol{\Sigma}_0) = (\hat{\boldsymbol{\beta}}_p^{**}, \hat{\boldsymbol{\beta}}_m^{**}, \hat{\boldsymbol{\Sigma}}^{**})$ and return to step (1). Otherwise, stop and define $(\hat{\boldsymbol{\beta}}_p, \hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\Sigma}}, \mathcal{L}) = (\hat{\boldsymbol{\beta}}_p^*, \hat{\boldsymbol{\beta}}_m^*, \hat{\boldsymbol{\Sigma}}^*, \mathcal{L}^*)$.

While objective function (9) is smooth enough for derivative-based quasi-Newton methods to function properly, the derivative-free pattern-search algorithms have the virtue of searching a broad array of points in the neighborhood of a candidate solution, which is useful in case local optima exist near the global optimum. Each one of the derivative-free methods differs by how it generates a set of local test points, so using all four ensures good coverage of the region around a candidate solution. In Step 1, ending with one last quasi-Newton solver provides a final Hessian matrix, which can be used for standard errors.

The process as described is a Gauss-Seidel approach to solving for a fixed point, but it can be sped up considerably by stopping in the middle of Step (2b) and proceeding on to Step (2c), if the researcher finds a non-trivial improvement before traversing all restart vectors. This adjustment brings it more in line with a Gauss-Jacobi fixed point method. Computation time averaged 12-36 hours per Tobit specification (Windows computer, 12$^{\text{th}}$ generation Intel processor (18 physical cores, 24 logical), 64GB RAM, MATLAB version 2022b).

Appendix D. Determinants of Initial Mathematics Proficiency

Although our ultimate interest is structural skill production technology—i.e., how home study activity (T_i, A_i) maps into skill gains $\Delta\kappa_i$—here we explore reduced-form determinants of initial skill, modeled as a Cobb-Douglas function with student UH traits θ_{pi} and θ_{mi} as the primary inputs:[9]

$$\kappa_i^{pre} = TFP_i \times \theta_{pi}^{\alpha_{pi}} \times \theta_{mi}^{\alpha_{mi}} \times \epsilon_i, \quad TFP_i > 0, \ \ \alpha_{pi} < 0, \ \ \alpha_{mi} < 0, \tag{15}$$

where TFP_i and production shares $(\alpha_{pi}, \alpha_{mi})$ are functions of covariates

$$\log(TFP_i) = \boldsymbol{X}_i\boldsymbol{\alpha}_0, \quad \alpha_{pi} = \boldsymbol{X}_i\boldsymbol{\alpha}_p, \quad \text{and} \quad \alpha_{mi} = \boldsymbol{X}_i\boldsymbol{\alpha}_m, \tag{16}$$

with $\boldsymbol{X}_i$ being a conformable $1\times(K+1)$ vector of student contextual factors, and $\boldsymbol{\alpha}_j$ being a $(K+1)\times 1$ parameter vector, $j = 0, p, m$. While equation (15) has a reduced-form interpretation—since it does not directly include endogenous investment from prior periods, but proxies for it using UH traits which drive endogenous investment—it provides a closer analog to previous VA models (discussed in Section 5.1.5) and previous empirical specifications of skill technology (see footnotes 4 and 5).

D.0.1. *Estimating the model.* Substituting (16) into (15), the initial skill model is equivalent to a regression of $\log(\kappa_i^{pre})$ on $\log(\theta_{pi})$, $\log(\theta_{mi})$, and $\boldsymbol{X}_i$, with pair-wise interactions:

$$\log(\kappa_i^{pre}) = \boldsymbol{X}_i\boldsymbol{\alpha}_0 + \boldsymbol{X}_i\boldsymbol{\alpha}_p\log(\theta_{pi}) + \boldsymbol{X}_i\boldsymbol{\alpha}_m\log(\theta_{mi}) + \log(\epsilon_i). \tag{17}$$

In our full specification, the covariate vector $\boldsymbol{X}_i$ contains a constant and all controls used in Table 6, *plus* parent survey controls (*Involved Parent*, *BigFam*, *Middle Child*, and *Youngest Child*). While the parent survey controls were not included in Table 6 due to numerical stability issues, this reduced-form model provides a suggestive robustness check on our school VA estimates there, by investigating whether the estimated relationship between skill stock and school dummies is robust to inclusion of additional household controls. Each factor may have a direct impact (through the TFP terms $\boldsymbol{X}_i\boldsymbol{\alpha}_0$), and an indirect impact (through the slope terms $\boldsymbol{X}_i\boldsymbol{\alpha}_p$ and $\boldsymbol{X}_i\boldsymbol{\alpha}_m$). In order to interpret reduced-form school VA effects causally, we require the following:

Assumption 8. $E[\boldsymbol{X}_i^\top\log(\epsilon_i)|\theta_{pi}, \theta_{mi}] = \boldsymbol{0}$.

Assumption 8 is weaker than analogous conditions commonly imposed in the prior literature, as it requires shocks to be exogenous, *conditional on UH* in productivity and motivation. It also highlights the advantages of our school VA identification strategy. Among our analyses in Sections 5 and 6, (17) is most comparable to the standard VA framework (equation (10)), but with three important changes. First, we can bring student ability ξ_i out of the error term and explicitly control for its influence on the outcome when measuring impacts of school dummies. Second, we need not assume latent ability is a single index; rather, $\xi_i = (\theta_{pi}, \theta_{mi})$, with different student traits having potentially distinct effects on the outcome. Third, we can capture the direct impact of school quality through its contribution to the $\boldsymbol{X}_i\boldsymbol{\alpha}_0$ (TFP) term, while allowing for heterogeneous treatment effects, or interactions between school quality and UH through the $\boldsymbol{X}_i\boldsymbol{\alpha}_p$ and $\boldsymbol{X}_i\boldsymbol{\alpha}_m$ (production share) terms.

[9]Recall that θ_p and θ_m are both inversely related to efficiency and motivation. Therefore, when a production share is larger in the *negative* direction, that is a *positive* thing for skill development.

Table OS.3. INITIAL MATH PROFICIENCY (Cobb-Douglas)

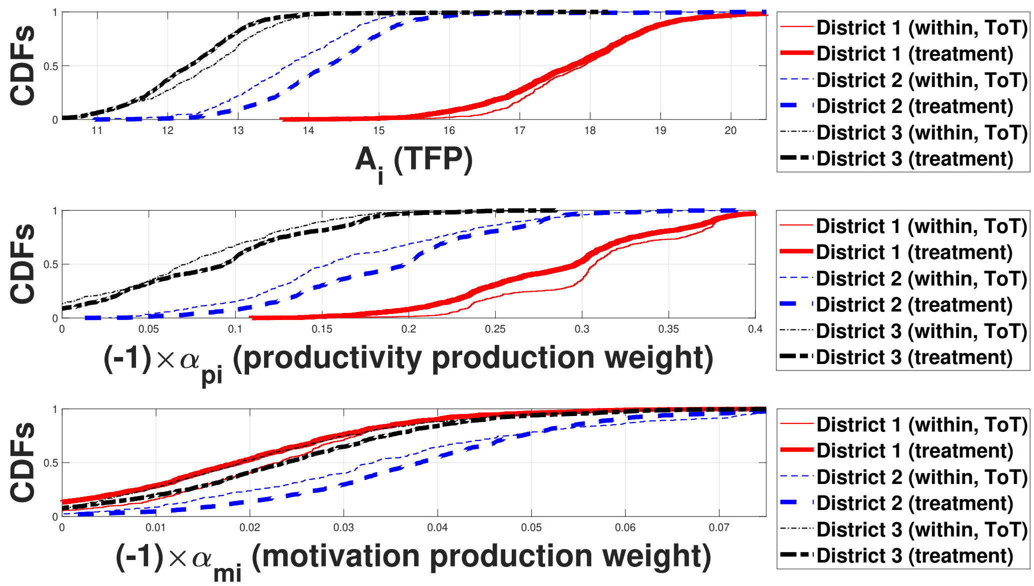
SPEC: **DEP VAR: log $(36 \times S^{pre}_{EB})$**	**(1) (Mean; SD)**	**(2) (Mean; SD)**	**(3) (Mean; SD)**	**(4) (Mean; SD)**	**(5) (Mean; SD)**
$(\log(\hat{TFP}_i))$	(2.849; 0)	(2.738; 0.173)	(2.728; 0.159)	(2.718; 0.173)	(2.730; 0.190)
θ_p **Production Share** $(\hat{\alpha}_{pi})$	(−0.292; 0)	(−0.230; 0.111)	(−0.216; 0.111)	(−0.208; 0.118)	(−0.201; 0.117)
θ_m **Production Share** $(\hat{\alpha}_{mi})$	(−0.020; 0)	(−0.025; 0.007)	(−0.024; 0.014)	(−0.025; 0.018)	(−0.022; 0.019)
	Avg SD Effect	**Avg SD Effect**	**Avg SD Effect**	**Avg SD Effect**	**Avg SD Effect**
log(*TFP*)	N/A	0.508***	0.466***	0.509***	0.558***
(joint p-value)		$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$
log(θ_p)	-0.606***	-0.478***	-0.448***	-0.432***	-0.406***
(joint p-value)	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$
log(θ_m)	-0.072***	-0.091***	-0.089***	-0.091***	-0.092***
(joint p-value)	(9.0×10^{-5})	(1.9×10^{-5})	(2.0×10^{-6})	(7.5×10^{-4})	(4.5×10^{-5})
Grade 5 $(\hat{\alpha}_{03}, \hat{\alpha}_{p3}, \hat{\alpha}_{m3})$	—	—	-0.137***	-0.129***	-0.146***
(joint p-value)			(1.3×10^{-8})	(2.2×10^{-9})	(2.0×10^{-8})
SCHOOL EFFECTS (District 1 omitted):					
Joint P-Value, All Terms ($df = 6$):		$(< 10^{-16})$	(8.7×10^{-10})	(8.3×10^{-6})	(9.6×10^{-5})
Joint P-Value, TFP Only ($df = 2$):		(9.9×10^{-10})	(2.0×10^{-4})	*(0.021)*	*(0.050)*
Joint P-Value, Slopes Only ($df = 4$):		(1.3×10^{-15})	*(0.064)*	*(0.081)*	*(0.091)*
District 2 $(\hat{\alpha}_{01}, \hat{\alpha}_{p1}, \hat{\alpha}_{m1})$	—	-0.264***	-0.220***	-0.241***	-0.167***
(joint p-value)		$(< 10^{-16})$	(8.5×10^{-7})	(2.6×10^{-5})	*(0.002)*
District 3 $(\hat{\alpha}_{02}, \hat{\alpha}_{p2}, \hat{\alpha}_{m2})$	—	-0.550***	-0.552***	-0.622***	-0.627***
(joint p-value)		$(< 10^{-16})$	(8.5×10^{-6})	(2.2×10^{-4})	(7.9×10^{-5})
Race/Gender (9)	no	no	YES***	YES***	YES***
Nbhd-SES Cntrls (6)	no	no	no	YES	YES
Home Acad Support (3)	no	no	no	YES	YES
Home Connectivity (9)	no	no	no	YES*	YES**
Prnt Survey Cntrls (12)	no	no	no	no	YES***
#Parameters	3	9	21	39	51
N	1,676	1,676	1,676	1,676	1,676
R^2	0.329	0.389	0.409	0.416	0.427

Notes: **Avg SD Effect** measures total impact of a variable through both TFP and production shares. It is the mean impact, in standard deviations of the outcome ($log(36 \times S^{pre})$), from switching a binary variable value from 0 to 1 (all else fixed), or increasing a continuous variable value by one standard deviation (all else fixed). Reported *joint p-values* are for the joint exclusion of all terms involving a given control from the model. Significance at the 99%, 95% and 90% levels are denoted by ***, **, and *, respectively. Stars on "YES/no" entries indicate joint significance.

D.0.2. *Empirical Results.* As outlined in Section 6.2.1, we compute EB forecasts of the left-hand-side variable, $\log(36 \times S^{pre}_{EBi})$ for efficiency, and we estimate equation (17) via FGLS, with heteroskedasticity-robust standard errors. Table OS.3 presents estimates of equation (17). For ease of interpretation, the table reports *Average SD Effects*, defined as mean induced shift in standard deviations of the outcome ($log(36 \times \kappa^{pre})$), from an increase in a continuous variable of one standard deviation, or a 0-to-1 change for binary controls. SD effects encapsulate influence through all channels, direct and indirect. While θ_p and θ_m are both significant determinants of initial math skill, productivity θ_p plays a dominant role with a SD effect that is roughly 4.5 times as big in specifications (4) and (5).

The second insight from Table OS.3 is evidence that school quality influences skill production technology in important ways. Magnitudes of the school district effects strongly conform to the pattern one might suspect from the suggestive evidence in Table OS.4: switching from District 1 (the high-resource district) to District 2 (the middling school district) or District 3

Figure OS.1. Cobb-Douglas Parameter CDFs by School



Notes: Since θ_p (θ_m) is inversely related to productivity (motivation), production share α_{pi} (α_{mi}) is typically *negative*, and lower two panels multiply it by -1 for ease of interpretation. Thin lines are CDFs for *students actually enrolled* in a given district (treatment on treated), and thick lines are general treatment effect CDFs for *all students*.

(the low-resource school district) entails substantially less effective skill production technology. We also see evidence of heterogeneous school treatment effects: school VA differences impact both TFP *and* the slopes of the production technology.[10] Figure OS.1 illustrates this idea with CDF-plots of i-specific production parameters. Finally, we again find evidence of decreasing returns to scale production technology in the sense that $-(\alpha_{pi}+\alpha_{mi})$ is well below the constant-returns benchmark of 1 for all i in the sample.

Given the similar identification strategy as in Section 6, the similarity of these results to Table 6 and the remarkable stability of school VA estimates in Table OS.3—including in specification **(5)** where additional parent-survey controls are added and are statistically significant—bolsters the causal interpretation of our school VA estimates in Table 6, suggesting they are not an artifact of omitted variable bias.

Appendix E. Supplemental Tables and Figures

[10] Table OS.3 specification (3) appears to somewhat favor TFP as the driver of school effects, though it becomes statistically harder to distinguish the two as more covariates are added in specifications (4) and (5).

Table OS.4. SCHOOL DISTRICT CHARACTERISTICS, AY2013-14

Variable	**STATE OF ILLINOIS**	**DISTRICT 1**	**DISTRICT 2**	**DISTRICT 3**†
FINANCES				
% Revenue from Local Property Tax	61.7%	85%	70%	35%
Operating Budget Per Pupil	$12,521	$14,500	$12,500	$13,500
% Spending on Instruction	48.7%	52%	48%	48%
FACULTY				
Avg. Administrator Salary	$100,720	$130,000	$105,000	$100,000
Avg. Teacher Salary	$62,609	$75,000	$60,000	$60,000
% Teachers w/Master's & Above:	61.1%	80%	65%	55%
Pupil-Teacher Ratio:	18.5	17	16	17
Pupil-Administrator Ratio:	173.3	210	140	130
STUDENT POPULATION & OUTCOMES				
% Low Income:	54.2%	0%	50%	90%
% Limited English Proficient:	10.3%	0%	0%	20%
% Meeting/Exceeding Expectations on State Math Exam (AY2014-15):	27.1%	60%	30%	10%

Notes: Data retrieved from the Illinois District Report Cards archive, 2015. District-specific numbers are rounded to preserve anonymity. **%Revenue from Local Property Tax** is rounded to the nearest 5 pp. **Operating Budget Per Pupil** is rounded to the nearest $500. **%Spending on Instruction** is rounded to the nearest 2 pp. **Avg. Teacher Salary** and **Avg. Administrator Salary** are rounded to the nearest $5K. **%Teachers with Master's & Above** is rounded to the nearest 5 pp. **Pupil-Teacher Ratio** is rounded to the nearest full number. **Pupil-Administrator Ratio** is rounded to the nearest 10. **%Low Income** is rounded to the nearest 10 pp and primarily represents students who are either from families receiving public aid or are eligible to receive free or reduced-price lunches. **%Limited English Proficient** is rounded to the nearest 10 pp. **%Meeting Expectations** is a measure adoptied by the Illinois State Board of Education for school performance. It roughly measures the fraction of a school's student body that is projected to be college-bound after graduation from high school. This measure is rounded to the nearest 10 pp and represents the average percentage across $5^{\text{th}}$and 6^{th}grades.

†**DISTRICT 3** had a relatively high operating budget per-pupil, but this is a misleading indicator of resources per student devoted to the core curriculum of math, science, social studies, and English. Like many districts serving less affluent communities, it receives additional state funding, but also faces additional expenses including social workers, meal subsidies, English as a second language programming, and non-instructional support programs.

Table OS.5. BALANCE TABLE

TREATMENT	**FEMALE**	**HISPANIC**	**Black**	**ASIAN**	**GRADE-5**	**PRE-TEST**	**#ASSIGNED SUBJECTS**
CONTRACT 1:	0.0005	-0.0054	0.0003	0.0032	-0.0014	-0.0021	557
(p-val)	*(0.99)*	*(0.82)*	*(0.99)*	*(0.90)*	*(0.95)*	*(0.93)*	
CONTRACT 2:	-0.0009	0.0024	-0.0048	0.0026	0.0001	0.0067	559
(p-val)	*(0.97)*	*(0.92)*	*(0.84)*	*(0.92)*	*(1.00)*	*(0.78)*	
CONTRACT 3:	-0.0009	0.0024	-0.0048	0.0026	0.0001	0.0067	560
(p-val)	*(0.97)*	*(0.92)*	*(0.84)*	*(0.92)*	*(1.00)*	*(0.78)*	

Notes: This table displays correlations between treatment assignment and the demographic and academic variables that were used for randomization. Treatment assignment randomization used balancing on gender, race, grade-level cohort, and pre-test score (via stratification). P-values (for the null hypothesis of zero correlation) are listed in parentheses.

FIGURE OS.2. Conditionally Heteroskedastic Work-Time Shocks

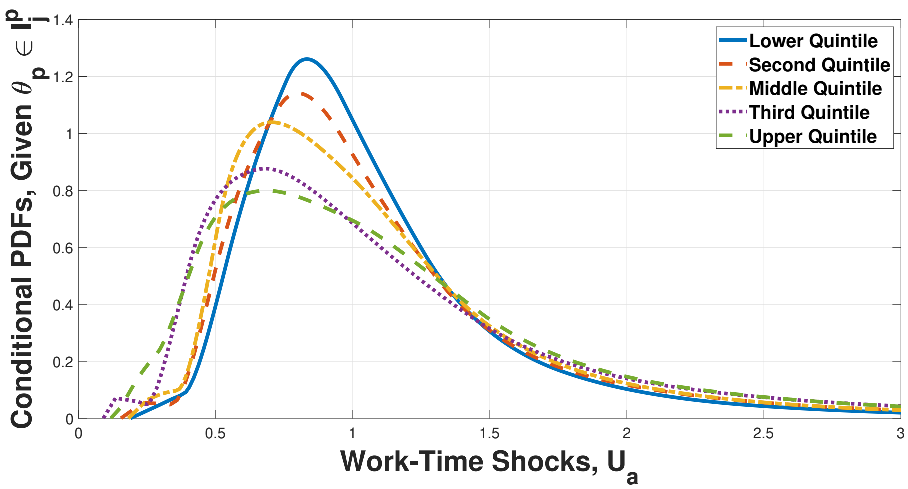


FIGURE OS.3. CDF Smoothing and Upper Tail Extrapolation