NBER WORKING PAPER SERIES

DISENTANGLING MOTIVATION AND STUDY PRODUCTIVITY AS DRIVERS OF ADOLESCENT HUMAN CAPITAL FORMATION: EVIDENCE FROM A FIELD EXPERIMENT AND STRUCTURAL ANALYSIS

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

Working Paper 27995
http://www.nber.org/papers/w27995

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020, Revised January 2025

Note: previous versions of this paper circulated under the title, "Productivity Versus Motivation in Adolescent Human Capital Production: Evidence from a Structurally-Motivated Field Experiment" We thank James Heckman and four anonymous referees for feedback that considerably improved this paper. Greg Sun, Nicholas Buchholz, Barton Hamilton, Stephen Ryan, Ismael Mourifié, Caroline Hoxby, Chris Taber, Jeffrey Smith, Samuel Purdy, Mary Mooney, Felix Tintelnot, Aloysius Siow, Angela Duckworth, Joseph L Mullins, Martin Luccioni, and Rob Clark also provided particularly helpful conversations regarding the content or exposition of this paper. Seminar participants at the University of Pennsylvania, the University of Chicago, the University of Wisconsin–Madison, Washington University in St. Louis, Queen's University, University of Toronto, the NBER Summer Education Meetings, and several other conferences and workshops provided useful feedback and suggestions. This project would not have been possible without a team of incredibly talented, dedicated, energetic, and tireless research staff, including: Debbie Blair, Edie Dobrez, Matthew Epps, Janaya Gripper, Clark Halliday, Allanah Hoefler, Justin Holz, Kristen Jones, No'am Keesom, Tova Levin, Claire Mackevicius, Wendy Pitcock, Joseph Seidel, Kristen Troutman, Andrew "Rusty" Simon, and Diana Smith. Finally, we are indebted to an army of research assistants 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, who were essential for executing our intricate experimental program. We are deeply indebted to the anonymous school administrators and teachers at our three partner school districts who generously went the extra mile to participate in this study. We also express our gratitude for the extensive discussions with Ariadne Merchant, Daphne Hickman, Morgan Hickman, Lydia Scholle-Cotton, and Nicholas Merchant. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Disentangling Motivation and Study Productivity as Drivers of Adolescent Human Capital Formation: Evidence from a Field Experiment and Structural Analysis

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

NBER Working Paper No. 27995
October 2020, Revised January 2025
JEL No. C93, I21, I24, I25, I26, J01, J24, O38

ABSTRACT

We estimate a structural model of endogenous short-run human capital investment focusing on a learner's leisure-study choices, influenced by external costs/benefits and two key internal factors: learning productivity and willingness to engage in study activity. Our novel self-investment framework rigorously quantifies models of learning that have existed in the psychology literature for decades. Our identification strategy combines panel data and study-incentive variation to point identify student-level parameters. Empirically, we find that idiosyncratic productivity and motivation traits are uncorrelated, and that low productivity is the stronger predictor of academic struggles, not low motivation. We investigate the influence of external factors on student learning and find that school quality affects it through 3 channels: augmenting productivity, augmenting skill production TFP, and by altering the mapping between learning activity and permanent skill gains.

Christopher Cotton
Queen's University
Department of Economics
Kingston, Ontario K7L 3N6
cc159@queensu.ca

Brent R. Hickman
Olin Business School
Washington University in Saint Louis
One Brookings Drive, Campus Box #1133
Saint Louis, MO 63130
hickmanbr@gmail.com

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and Australian National University
and also NBER
jlist@uchicago.edu

Joseph Price
Department of Economics
Brigham Young University
2129 WVB
Provo, UT 84602
and NBER
joseph_price@byu.edu

Sutanuka Roy
The Australian National University
HW Arndt Building 25A
Kingsley Pl
Acton ACT 2601
Australia
sutanuka.roy5@gmail.com

1. Introduction

Suppose that Anthony, a 6$^{\text{th}}$-grade student, does not regularly complete his mathematics homework and performs poorly on exams. Seeing this, one might conclude that Anthony's preferences are at fault: he either lacks motivation for math study or places low value on academic achievement. If so, it may be possible to improve Anthony's outcome by providing incentives or information on the returns from schooling to him or his parents. However, low motivation is not the only possible explanation for Anthony's observed choices and outcomes. He may be very willing to put in work, but if he lacks foundational skills, adequate study support, or high-quality school instruction, then spending even large amounts of time on math may be inadequate to raise his grades. Thus, Anthony may rationally withdraw from study activity, despite a basic willingness to work. In this case, a very different intervention is needed, and merely nudging Anthony's beliefs or incentives would be ineffective.

The challenge for policymakers, educators, and researchers stems from inferring whether low-performing students lack motivation—willingness to allocate a fixed quantity of time to study—or whether they struggle with academic productivity—ability to convert time inputs into completed assignments or improved exam scores. Typical observational datasets focus on outcomes (e.g., grades/exam scores), but include only coarse student time inputs or none at all, meaning both explanations—low motivation and low productivity—are observationally equivalent.[1] Moreover, observational data cannot identify a mapping between day-to-day learning activity and incremental skill gains. These inferential problems are a formidable barrier to understanding student outcomes, both at the individual and group level.

To overcome these challenges, we execute a structurally-motivated field experiment involving 1,676 adolescent math students in partnership with their schools, teachers, and parents.[2] The setting of a natural field experiment with automated tracking of web-based learning activities allows us to observe (*i*) how study effort responds to incentives, (*ii*) how effectively students convert time inputs into completed assignments, and (*iii*) how learning activity maps into measurable skill gains. We directly quantify structural labor-supply elasticity and study-time productivity for each individual student, shedding new light on the root causes of low academic performance, and its link to a learner's rational day-to-day choices. Furthermore, by running the experiment across a diverse set of students and school districts, we can explore how motivation and productivity are influenced by contextual factors such as prior math skill, demographics, and school quality. Our field experiment produces a rich set of student-level variables and varying incentives that are well-tailored to solve various empirical confounds present in observational data from educational settings.

[1] The most closely related papers are Cunha, Heckman, and Schennach (2010) and Agostinelli and Wiswall (2022), which use NLSY data at two-year intervals, and Del Boca, Flinn, Verriest, and Wiswall (2019), which uses PSID data including child time-use variables. Neither dataset (or any others we know of) includes time-use data and exogenous incentive variation needed to directly quantify children's labor-supply elasticities.

[2] This study adhered to strict standards of research ethics; see Section 3.2 for further discussion.

To analyze our field-experimental data, we propose a novel quantitative framework for day-to-day labor-supply choice by learners. Leveraging the psychology literature (e.g., Carroll (1963); Eccles and Wigfield (2002)), we model each student as having two idiosyncratic characteristics that govern productivity and the opportunity cost of time. A child's utility costs of spending study time are convex, meaning that she becomes increasingly less willing to continue work as her total time allotment to studying grows. This model feature admits various natural interpretations, including physical/mental exhaustion, or marginal value of non-study time rising as other activities become increasingly crowded out.

A central implication of this framework is that design of academic compensation schemes drive learner behavior in important ways. Virtually all incentives tied to learning are "piece-rate," where individuals are rewarded for outputs rather than for time inputs. Continuing the thought experiment, if Anthony and Joseph both earn an "A" in a math course, they both see the same improvement in their chances for job/college applications, even if Joseph required only a fraction of Anthony's study time. Within our model, while a child's motivation characteristic alone determines willingness to spend an hour studying, under piece-rate incentives his productivity characteristic will also play a central role in his decision of whether to spend *enough time* to achieve at a high level. A child with high motivation (i.e., low cost of studying an extra hour) may still choose to complete few assignments if an hour of his study time is sufficiently unproductive for reaping the rewards of achievement.

Our approach to empirically modeling education is novel for two important reasons. First, we focus on aspects of adolescent human capital—day-to-day rational leisure-study choice—that have not been thoroughly studied in the literature. Second, we depart from typical methods and models because our novel data collection procedure opens a window into learner behavior not previously possible with observational data. A comparison to other empirical work on childhood skill development is informative for understanding our contribution. A branch of this literature focuses on parental investment in child human capital and/or policy interventions such as financial resources or incentives for parents; e.g., Cunha et al. (2010) Del Boca, Flinn, and Wiswall (2014), Fryer, Levitt, and List (2015), Chetty, Hendren, and Katz (2016), Del Boca et al. (2019), Agostinelli and Wiswall (2022), Gayle, Golan, and Soytas (2022). Another branch of the literature focuses on the importance of schooling-input quality/quantity; e.g., Hanushek (2020), Dobbie and Fryer (2011), Cullen, Levitt, Robertson, and Sadoff (2013), Chetty, Friedman, and Rockoff (2014), Fryer (2017), Guryan et al. (2021), Ahn, Aucejo, and James (2022), Fryer, Levitt, List, and Sadoff (2022), Luccioni (2023)). These are important topics deserving of scientific attention, but parents and schools are only part of the parent-child-school partnership that shapes adolescent human capital development.

Much less is known about how a child's own decisions affect skill development. Indeed, one could argue that learner choices are the one truly indispensable input. Low school quality may be offset by intensive parental inputs, while many public programs are geared toward

partially offsetting low parental inputs in a child's education—e.g., universal pre-K, free and reduced-price lunch, and after-school programs. If parental and school educational inputs are both lacking, personal effort of the learner may even substitute for both. Prominent historical examples include Alexander Hamilton, Frederick Douglass, and Abraham Lincoln, while contemporary examples include authors Jeanette Walls (2017) and Tara Westover (2018). However, if the learner himself/herself is unwilling or incapacitated from contributing to the parent-school-child partnership, it is difficult to imagine a viable compensating factor in human capital production. Our research is designed to provide a novel and detailed focus on the learner's rational decision process of *self-investment*.

A newer branch of the literature experimentally studies children's responses to incentives as a means of spurring academic improvement; e.g., Fryer (2011), Fryer (2016) Levitt, List, and Sadoff (2016), Burgess, Metcalfe, and Sadoff (2016), and C. Cotton, Hickman, and Price (2022). While compelling as a growing body of evidence, Fryer (2011) articulated the limitation of this standard approach: "*we urge the reader to interpret any results as specific to these [experimentally tested] incentive schemes and refrain from drawing more general conclusions.*" Furthermore, with the sole exception of C. Cotton et al. (2022), these studies do not produce real-time data on children's day-to-day behavior changes in response to incentives, or data on how these altered behaviors map into incremental academic improvement.

Our study and primary research questions are geared toward developing generalizable insights on student choices, rather than just the impact of a specific incentive scheme or policy. How responsive are a child's labor-supply choices to a given level of enticement toward math work? How productive is their time when they do spend it studying? How heterogeneous are adolescents in their study-time productivity and willingness to substitute time from their outside options toward schoolwork? What portion of this variation is attributable to observable external differences (e.g., family background, outside time-use options, school quality, socioeconomics, etc.) versus heterogeneity that is idiosyncratic to the child? How does this heterogeneity shape adolescent skill-production technology, and what role(s) does school quality play? And finally, how do a child's motivation and productivity traits interact to produce choices and outcomes, under piece-rate compensation schemes?

We find strong evidence that a large fraction of struggling students are not less motivated than their high-performing peers, but rather, they find difficulty in converting time inputs into success. We also document surprisingly large productivity heterogeneity among academic high-achievers. Our structural model also points to labor-supply cost convexity as a significant factor in driving student choices. For the median student, we estimate that a doubling of the child's daily time commitment to math study causes monetized utility costs to rise by a factor of more than three.

We also investigate the extent to which contextual factors can explain productivity and motivation differentials across students. While much of this heterogenetiy appears idiosyncratic to the child—56% for productivity and 71% for motivation—in both cases, we are able

to identify predictors that play a meaningful role in enhancing (or detracting from) a child's academic traits. Due to the pooled-cohorts design of our field experiment, we can measure the year-on-year change in academic productivity and motivation: 5^{th}-graders require more time inputs to solve homework problems by 30% of a standard deviation, relative to 6^{th}-graders, while both age groups are comparably willing to allocate time to math and away from outside uses. One of our main empirical findings is that school quality is a substantial factor for augmenting academic productivity (i.e., rate of progress through homework tasks), sometimes having an effect larger than the age-cohort difference.

Turning to an investigation of skill production technology, we find that between overall study time and (quality-controlled) learning assignment completion, the latter is the primary driver of incremental skill gains. Student productivity traits not only spur higher assignment completion, but more productive students tend to also convert a fixed volume of math learning tasks into permanent skill gains as well. Estimates also point toward a decreasing returns-to-scale learning technology: math score gains of a fixed margin become harder to accomplish as a child's baseline proficiency rises. We also find evidence that school quality plays a significant role in shaping human capital production technology through two channels: it augments TFP and also increases the rate at which a fixed volume of completed learning tasks are converted into permanent math skill gains.

Our work combines two literatures on estimation of skill formation technology (e.g., Cunha et al. (2010); Agostinelli and Wiswall (2022)) and school value-added (e.g., Koedel, Mihaly, and Rockoff (2015); Abdulkadiroglu, Pathak, Schellenberg, and Walters (2020)). Our experimental data enable us to produce novel contributions to the value-added literature in several ways. First, an important advantage of our research design is that identification of the structural model does not require information on typical endogenous observables like exam scores or school assignment. Instead, we pin down mutli-dimensional unobserved student traits with a combination of real-time data on home-study activity and experimental incentive variation. This facilitates secondary analyses, based on our structural estimates, where we can directly control for the canonical problem of selection on unobservables when estimating school effects on skill formation technology. Second, the previous literature on school value-added has typically conceptualized latent student ability as a single index (see survey by Koedel et al. (2015)), whereas our experimental design enables us to quantify two distinct, outcome-relevant dimensions of student ability—productivity and motivation—with distinct impacts on the process of skill formation.[3] Third, and related to the second, our experimental observables, which include multiple test scores and fine-grained data on interim home-study activity, allow us to expand upon traditional measures of value-added to explore mechanisms through which school quality may operate. Specifically, we allow for the possibility that better schools improve outcomes (i) by augmenting a child's study productivity,

[3]These traits are similar in spirit to the cognitive and non-cognitive skills studied by Cunha et al. (2010).

(*ii*) by augmenting motivation, (*iii*) by augmenting skill production "TFP," (i.e., skill gains arising independently of home study), and (*iv*) by increasing the rate at which homework accomplishment translates into permanent skill gains. Our estimates suggest that three of these four channels ((*i*), (*iii*), and (*iv*)) play a meaningful role in shaping adolescent academic choices and outcomes. We also find evidence suggestive of a final, indirect value-added channel. More productive learners not only choose to do more homework, but they tend to glean more permanent skill gains from a fixed volume of homework; to the extent that school quality strengthens productivity, it may also strengthen this mechanism as well.

Our study highlights the importance of one's learning environment in shaping both rational academic choices and outcomes. Even when controlling for various individual and family characteristics, we find that those who attend schools in more affluent districts have higher learning productivity and are better able to convert study effort into measured skill gains. Even highly-motivated and productive students tend to perform significantly better, all else equal, when they attend a high-performing school than a low-performing school. Moreover, decreasing returns-to-scale skill production technology suggests the social value of investing in learning environments is highest among students who lack quality educational resources.

Our paper is structured as follows. Section 2 outlines the theoretical framework that underpins our research design. Section 3 describes the field experiment and Section 4 presents identification and estimation of the structural primitives. Section 5 decomposes structural student type parameters by a host of observable factors. Section 6 performs secondary analyses on math skill production functions and Section 7 concludes. An Online Appendix contains additional technical details, graphs, tables, and details of our methodology.

2. A Theoretical Framework for Learner Self-Investment

We propose a quantitative model of rational learner choice in the spirit of qualitative frameworks from psychology known as "mastery theory" (e.g., Carroll, 1963) and "expectancy-value theory" (e.g., Eccles et al., 1983; Wigfield, 1994; Eccles & Wigfield, 2002; Wang & Degol, 2013). The psychology literature frames skill acquisition as a sequence of learning tasks: an algebra student attends class, is assigned homework problems, chooses how much of each assignment to complete, and then iterates the process each day. At the end of the course, the child's measured algebra competency is determined by the cumulative volume of learning assignments he completed. Thus, human capital attainment hinges on a series of high-frequency decisions made by a child on a day-to-day basis, over a course of weeks or months. Our goal is to formally quantify the structural elements of this decision process. Core model primitives include a 2-dimensional vector of learner traits that shape the mapping between effort and rewards. While unobserved heterogeneity and leisure-study decicions are the core of our empirical exercise and the sole focus of the first half of the paper, Sections 5 and 6 show how the model we develop can open additional avenues for policy-relevant analysis,

such as decomposing environmental and idiosyncratic factors that drive education outcomes, and solving endogeneity problems that plague estimation of school value added.

2.1. **A Formal Model of Study-Leisure Choice.** Learning choices depend on (*i*) how easily/quickly a child can complete learning tasks, and (*ii*) her perceived value of success relative to the cost of effort. We refer to these components of incentives as *study-time productivity* and *motivation*. For each child, indexed by i, let $A_i \in \mathbb{N}$ denote the number of learning tasks that i completes within a fixed period of time. The precise definition of a "time period" is not crucial, provided it is short enough that a child's decision-relevant characteristics can be thought of as fixed and stable within a period—e.g., a week, a month, or a semester. Our model is one of within-period decisions, with opportunity costs of allocating a certain fraction of a finite time endowment to math study.

Each individual task, chronologically indexed by $a_i = 1, 2, \ldots, A_i$, is a discrete unit of work; e.g., a single math problem, a block of problems, or an entire assignment. Completion of these learning activities builds skill proficiency, which is of ultimate interest to policy-makers. We study the mapping between cumulative work and measured proficiency gains in Section 6, but at present we focus solely on within-period rational study-time choice, $T_i(A_i)$.

Definition 2.1. *(Inverse) Study-time productivity*, denoted $\theta_{pi} > 0$, governs the rate at which child i is able to complete learning tasks.

Study-time productivity is inversely proportional to θ_p, but for ease of discussion we refer to it simply as "productivity".[4] The mapping $T_i(A_i) : \mathbb{N} \to \mathbb{R}_+$. is stochastic, with total time commitment being an aggregate of study times across individual completed learning tasks:

$$T_i(A_i) \equiv \sum_{a_i=1}^{A_i} \tau_i(a_i; \theta_{pi}); \quad \tau_i(a_i; \theta_{pi}) \equiv \theta_{pi} \times \tau_0 \times \tau_1^{\mathbb{1}(a_i=1)} \times a_i^{-\varphi} \times U_{a_i}, \quad \tau_0,\ \tau_1,\ U_{a_i} > 0. \tag{1}$$

Here, $\tau_i(a; \theta_{pi})$ is the time required for i to complete her a^{th} learning task. τ_0 is mean completion time across all students, τ_1 is a "startup cost," on the first task of the period, and the productivity fixed effect θ_{pi} scales mean completion time up or down. The term $a_i^{-\varphi}$ is a standard "experience curve" (Wright (1936)) whereby a student's rate of progress may grow (if $\varphi > 0$) or decay (if $\varphi < 0$) with in-period task volume. The experience curve term allows for short-run gains in productivity, despite θ_{pi} being fixed within-period. One concern is over the assumption that student productivity types are fixed within a period. Could students' productivity be updating in real-time as they traverse learning tasks? Ultimately this is an empirical question of the time frame over which one can appropriately apply a short-run model with stable latent types. However, including the learning curve term $a_i^{-\varphi}$ serves as a robustness check by allowing for short-run productivity growth (or decay). As a preview, we estimate very small short-run experience effects—a 5% time reduction after *doubling* output, or $\frac{\tau_i(2a_i;\theta_{pi})}{\tau_i(a_i;\theta_{pi})} \approx 0.95$—but large year-on-year shifts in θ_{pi} across 5^{th}- and 6^{th}-graders.

[4] In Section 6 we also allow for θ_{pi} to influence the rate at which work volume (T_i, A_i) is converted into permanent skill gains.

The transitory shock, U_{a_i}, is *iid* across tasks and represents unpredictable fluctuations in difficulty, mental state, distractions, etc. The element of randomness provides a realistic representation of the data: we observe substantial variation of within-student completion times across learning tasks (see Table 1). It also squares well with common academic experience: sometimes a dreaded math problem turns out to be easier than expected, and other times a learner's initial optimism melts away as a given exercise drags on. We use upper-case U to denote random outcomes and lower-case u for specific realizations. Subscripted "F" denotes exogenous distributions, while subscripted "G" indicates a distribution of some endogenous object. We assume the shock distribution is well-behaved in the following sense:

Assumption 1. *The production-time shock U_{a_i} follows (heteroskedastic) distribution $F_u(u_{a_i}|\theta_{pi})$ with continuous density f_u that is bounded away from zero on nonempty support $[\underline{u}, \overline{u}] \subset \mathbb{R}_+$.*

For our empirical study we focus specifically on math learning, though the model admits various interpretations about what A_i and T_i represent. They could represent general schoolwork, in which case the outside option for time encompasses all non-scholastic activity (e.g., sleep, chores, socializing, recreation, etc.). Alternatively, they could represent subject-specific inputs (e.g., math), in which case the set of outside options for time includes work on all other school subjects *and* non-school activities. In this case, a child faces outside incentives for all activities, including science homework, and diverting time towards math makes it more difficult to attain rewards for science achievement. Thus, we model the cost for individual students to substitute time away from the most profitable outside use (including homework in English, Science, etc.), and toward math learning instead.

Definition 2.2. *(Inverse) Motivation*, denoted $\theta_{mi} > 0$, indexes idiosyncratic labor-supply costs, or i's willingness to substitute a fixed quantity of time toward math activity.

Although willingness to spend time studying is inversely related to θ_m, for ease of discussion we often refer to it simply as "motivation." Student i's cost of spending T_i hours learning math is multiplicatively separable in her motivation type and a common labor-supply cost function: $C_i(T_i; \theta_{mi}) \equiv \theta_{mi} c(T_i)$. The separability assumption allows for a broad degree of flexibility in the form of the cost function, while aiding identification by imposing a homogeneity condition that Bodoh-Creed, Hickman, List, Muir, and Sun (2023) refer to as "rank-stability:" if agent i's effort level is at the p^{th} percentile under one incentive scheme, then i will remain at the p^{th} percentile under any affine incentive shift. Assumption 2 establishes regularity conditions that ensure a well-behaved leisure-study decision problem.

Assumption 2. *Costs are twice differentiable, $c'(t) > 0$, and $c''(t) > 0$ $\forall t \in \mathbb{R}_+$; marginal costs $c'(t)$ are unbounded, and we adopt normalizations $c(0) = 0$ and $c'(0) = 1$.*

Intuitively, the cost of allocating time to study rises as a child spends more time working. Likewise, i's cost levels and marginal costs for any $t > 0$ are increasing in the inverse motivation parameter: high θ_{mi} means that child i incurs relatively more dis-utility from an hour of

study. Cost convexity (positive second derivative) has an intuitive interpretation: marginal disutility of sacrificing outside options rises with one's total math-work time. That is, each additional hour studying math is more costly than the previous one, because either marginal utility of outside options rises as their consumption is increasingly crowded out, or because the direct marginal psychic cost of math effort is rising, or both.[5]

Our specification of costs does not derive from an explicit model of a child's complete time-allocation decisions on sleep/grooming/eating, school time, regular study, leisure, social time, scheduled extracurricular activities, chores, extracurricular website activity, etc. The advantages of this approach are two-fold. First, it provides a simple and tractable model of how a child trades off external incentives to study math versus utility from optimal outside time usage. Second, this approach has the potential to allow for some generality, as our specification of the cost function need not pre-suppose a specific model of all student time allocation choice.[6] Given that our cost function $C(t;\theta_{mi}) = \theta_{mi}c(t)$ does not derive from an explicit model, we interpret the parameter θ_{mi} as a "reduced-form" estimate.[7] In Section 5 we draw upon a wealth of student-level observables to empirically tease apart observable factors that contribute to student types. However, for the purpose of the structural model, θ_{mi} may encompass *either* intrinsic costs of effort, opportunity costs of time, innate characteristics of child i, environmental factors, or some mixture of these. A similar statement applies to θ_{pi} as well: in the common language of regression analysis, $(\theta_{pi}, \theta_{mi})$ is a two-dimensional fixed effect encompassing all factors of i or her life circumstances that are stable over the short run, and govern rate-of-progress and leisure-study tradeoffs, respectively.

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

Intuitively, piece-rate incentive schemes may encompass all external "carrots" and "sticks" presented to child i by her home, school, and community. Parents may inculcate in her a positive intrinsic value of achievement, or they may offer tangible rewards or punishments for completion or non-completion of work and achievement benchmarks. A child's school rewards her for homework completion with grades, and may further motivate her regular coursework via pre-announced exams, with cumulative grades determining her future education and career prospects. Additionally, businesses, organizations, and colleges may offer merit-based admissions, internships, or scholarships that improve expected flows of future monetary and psychic income from a desirable career (Becker (1993)).

[5]Unbounded marginal costs ensures finite study choices. E.g., if a period is a week, and t is hours, then one might naturally assume $c'(t) \to \infty$ as $t \to 168$. This limiting choice would entail 7 full days of uninterrupted math study, requiring extreme and physically dangerous levels of sleep and food deprivation.

[6]Agostinelli and Wiswall (2022) used a similar strategy to model parental investment in adolescent HC.

[7]One child may have more valuable outside uses of her time available, such as a new gaming system or a prolific friend network. Child i may simply incur larger psychic costs of exerting effort solving math problems, relative to j. Alternatively, j's home/school may engender different norms regarding the value of work.

2.1.1. *Optimal Choices.* In choosing optimal work (T_i, A_i), child i solves an optimal stopping problem by recursively comparing costs and benefits of an additional completed assignment, while accounting for randomness.[8] Given $(a_i{-}1)$ completed assignments, a student optimally determines the maximum time, $t^*_{a_i}$, she is willing to devote to the a_i^{th} task. Equation (1) implies success probability on task a_i, given t units of time input, is $F_s(t; a_i, \theta_{pi}) \equiv \Pr\big[\tau_i(a_i;\theta_{pi}) \leq t | a_i\big] = \Pr\big[U_{a_i} \leq \frac{t}{\theta_{pi}\tau_0\tau_1^{\mathbb{1}(a_i=1)}a_i^{-\varphi}}\big] = F_u\left(\frac{t}{\theta_{pi}\tau_0\tau_1^{\mathbb{1}(a_i=1)}a_i^{-\varphi}}\middle|\theta_{pi}\right)$, with first derivative denoted by $f_s(t;a_i,\theta_{pi})$. A learner's decision problem is defined by the Bellman equation,

$$\mathcal{V}(a_i{-}1, T_i(a_i{-}1)) = \max_{t\geq 0}\left\{F_s(t;a_i,\theta_{pi})\Big[\Big(\Pi_i(a_i){-}\Pi_i(a_i{-}1)\Big) + \mathcal{V}\Big(a_i, T_i(a_i{-}1)+t\Big)\Big] - \theta_{mi}\Big[c\Big(T_i(a_i{-}1)+t\Big){-}c\Big(T_i(a_i{-}1)\Big)\Big]\right\}. \quad (2)$$

The first term inside the curly brackets is payoffs from work, being success probability times the sum of immediate incremental payoffs $(\Pi_i(a_i) - \Pi_i(a_i{-}1))$ and continuation value

$$\mathcal{V}(a_i, T_i(a_i)) \equiv \max_{\tilde{t}}\left\{\mathrm{E}\left[\tilde{A}|\tilde{t}, a_i, \theta_{pi}\right]\Big(\Pi(\tilde{A}) - \Pi_i(a_i)\Big) - \theta_{mi}\Big[c(T_i(a_i)+\tilde{t}){-}c(T_i(a_i))\Big]\right\}, \quad (3)$$

where $\tilde{A}$ and $\tilde{t}$ are future tasks completed and future time worked *after completing the* a_i^{th} *learning task*. Note that a student retains the opportunity to reap rewards from future work only if she does not walk away during the a^{th} task attempt. This is an innocuous assumption, given that a_i merely represents a chronological index on work i chooses to complete. The last term inside the brackets is incremental costs from work time on the current task a_i.

Thus, optimal choice t^*_a is implicitly defined by the first-order condition[9]

$$f_s(t^*_{a_i}; a_i, \theta_{pi})\Big(\big[\Pi_i(a_i){-}\Pi_i(a_i{-}1)\big] + \mathcal{V}\Big(a_i, T_i(a_i{-}1){+}t^*_{a_i}\Big)\Big) + F_s(t^*_{a_i}; a_i, \theta_{pi})\mathcal{V}_2\Big(a_i, T_i(a_i{-}1){+}t^*_{a_i}\Big) = \theta_{mi}c'\left(T_i(a_i{-}1){+}t^*_{a_i}\right). \quad (4)$$

Intuitively, if she completes the a_i^{th} assignment in time $t < t^*_{a_i}$, then she pauses and re-optimizes the updated Bellman equation (2) for the $(a_i{+}1)^{\text{th}}$ learning task. Otherwise, if devoting time $t{<}t^*_{a_i}$ does not complete assignment a_i, she continues to work and her stopping time $t^*_{a_i}$ balances the expected marginal benefit of continuing work (including retention of future payoff opportunities) against the deterministic marginal cost. If she reaches $t^*_{a_i}$ without realizing her a_i^{th} success, she discontinues work for the remainder of the period, and $(T_i(A_i), A_i)$ are determined by her recursive optimal stopping point, where $A_i{=}(a_i{-}1)$.

In reality, $A_i{=}A_i\left(\theta_{pi}, \theta_{mi}, \Pi_i, \{u_{a_i}\}_{a_i=1}^{A_i+1}\right)$ and $T_i{=}T_i\left(\theta_{pi}, \theta_{mi}, \Pi_i, \{u_{a_i}\}_{a_i=1}^{A_i+1}\right)$ both depend not only on a child's characteristics and incentives, but also on a specific realized history of shocks U encountered along the way to her stopping point. For notational compactness we suppress this dependence, but note that conditional on a given $(\theta_{pi}, \theta_{mi}, \Pi_i)$ triple, there is a non-degenerate distribution of final within-period choices, $G_{ta}(T_i, A_i|\theta_{pi}, \theta_{mi}, \Pi_i)$.

Two key model predictions are relevant to model identification. First, if a child continues on to the $(a_i{+}1)^{\text{th}}$ task there is an important shift in her decision problem: previously her accrued cost baseline was $T_i(a_i{-}1)$, while now it is $T_i(a_i) > T_i(a_i{-}1)$. Thus, cost convexity implies

[8] We use the term "optimal stopping problem" in the sense of the statistical decision theory literature pioneered by Wald (1945), Arrow, Blackwell, and Girshick (1949), Snell (1952), and Chow and Robbins (1963).

[9] In equation (4), if the left-hand side is less than the right-hand side given $a_i{=}0$ and $T_i = 0$, then a corner solution obtains.

elevated marginal costs of continuing, and bounded marginal payoffs ($\Pi_i(a_i+1)-\Pi_i(a_i)\leq\overline{\pi}$) imply her maximal willingness to work eventually declines, or $t^*_{a_i+1}<t^*_{a_i}$. Second, the model predicts a monotone relationship between student types and actions. More precisely, the stochastic mapping from unobserved θ_{mi} to observed A_i, conditional on fixed θ_{pi}, exhibits a monotone likelihood ratio: reductions in labor-supply costs lead to first-order dominance shifts in a child's distribution of work volume, or $\theta_m<\theta'_m \Rightarrow G_a(a|\theta_p,\theta_m)\leq G_a(a|\theta_p,\theta'_m)$. Likewise, a similar monotone-likelihood-ratio property holds for the relationship between θ_{pi} and A_i (conditional on fixed θ_{mi}): $\theta_p<\theta'_p \Rightarrow G_a(a|\theta_p,\theta_m)\leq G_a(a|\theta'_p,\theta_m),\ a\in\mathbb{N}$.

2.2. **The Significance of Piece-Rate Incentives.** Piece-rate compensation—that is, $\Pi_i(\cdot)$ is a function of completed tasks, A_i, rather than time inputs, T_i—is the dominant form of incentive provision in academic settings. If two students, *Tabby* and *Jane*, both complete 9 out of 10 math assignments, and score 95% on the final exam, academic rewards do not distinguish between *Jane*'s 40-hour math commitment versus *Tabby*'s 20 hours. Both children receive the same grade as a result of their identical performance record and the same improved prospects for their desired college seat, scholarship, job, etc. This modelling choice is not only empirically relevant, but it also has profound implications for how incentives interact with a child's unobserved traits. If we consider a switch to time-based incentives, say $\widetilde{\Pi}_i(t)$, then (2) reduces to a simpler decision problem where a child's productivity θ_{pi} is irrelevant to their optimal study-leisure choice. By contrast, in the piece-rate decision problem (2), rational choice of time commitment depends not only on how averse a child is to spending an hour on math (i.e., θ_{mi}), but also on *how productive an hour of her time will be* (i.e., θ_{pi}) for reaping output-contingent rewards of work.

Our model of adolescent time-allocation therefore immediately calls into question prevailing wisdom behind labels that are often applied based on observed behaviors. For example, if *Jane* turns in only 50% of her assignments while *Tabby* completes all of them, many practitioners and researchers simply conclude that *Jane* is "unmotivated" for math study, while *Tabby* appears "well-motivated." While the model allows for this as a plausible interpretation, it is also equally plausible that Jane turns in less homework despite being *more* motivated than *Tabby* ($\theta_{m,Tabby}\geq\theta_{m,Jane}$) if she is sufficiently less productive ($\theta_{p,Tabby}<\theta_{p,Jane}$). Thus, both explanations are observationally equivalent given the single raw data point of *Tabby*'s and *Jane*'s study choices or academic outcome.

The model also calls into question various common and incorrect usages of the term "effort." In the example above many would say that *Tabby* put forth more "effort" since she completed more work. However, if *Tabby* is more than twice as productive as *Jane*, then *Jane* actually *spent more time* working on math to produce half as much output, and can be said to have exerted greater "effort" than *Tabby*. Our model highlights how multiple dimensions of agent unobserved heterogeneity may imply that there does not exist a one-to-one mapping between typically unobserved measures of true effort (e.g., time spent, personal costs incurred, etc.) and observable output (e.g., grades, homework, etc.).

2.3. **Discussion on Modelling Choices.** Before moving on, we briefly discuss some aspects of our modelling approach. First, our goal is to study short-run adolescent decision processes, so there are no formal "future periods" in the model, aside from the chronological indexing, "a," of learning tasks. However, cost convexity essentially performs the role of "discounting" expected utility on later units of work (i.e., $a_i' > a_i$), since the baseline marginal cost of zero additional effort on the next unit (a_i+1) rises with time spent on task a_i.

On a related note, the decision model is a non-stationary dynamic program because of cost convexity combined with history-dependence of the state variables $(a_i-1, T_i(a_i-1))$. The continuation value argmax in (4) updates continuously as time accrues on assignment a_i, and the distribution of payoffs on future work $\widetilde{a} > a_i$ is not known until the student finally completes task a_i.[10] The model is computationally taxing for the researcher, but it has several advantages, the most important being that it requires only simplistic thinking on the part of the adolescent decision-maker. The optimal stopping model merely assumes a child comprehends three basic pieces of information at each point in time: (i) her marginal incentives to complete another task, (ii) the variability of completion times for current and future tasks, and (iii) how taxed/exhausted she feels from previous work. In short, a child need only be aware of her current feeling and of her ability to continue productively.

An alternate model where children make a one-shot decision on achievement target A_i would be computationally simpler, but requires much stronger assumptions: learners must plan ahead and rigidly stick to their ex-ante study plan, regardless of whether they experience a string of especially good study-time shocks along the path to A_i. In contrast, if a child completes A_i assignments unexpectedly quickly, she would have incentive to work beyond the target A_i: her marginal cost $c'(T_i(A_i))$ would still be relatively low, allowing her to reap expected marginal profits. Similarly, a particularly bad string of shocks would motivate her to abandon work prematurely. Therefore, the simpler one-shot decision model would require student behavior that is not always ex-post rational. Thus, we adopt the optimal stopping model as a more defensible empirical framework for real-time leisure-study trade-offs.

3. A Field Experiment to Identify Student Motivation and Productivity

Observational equivalence between opposing explanations for observed behaviors means that the model of adolescent study-leisure choice is *not* identified from standard education data. This fact motivates our field-experimental design, carefully crafted to include multiple dimensions of student observables and exogenous variation unavailable in observational data. Our research design forms part of a nascent literature that employs field experimental methods for identifying structural primitives of an economic model, rather than to directly test hypotheses about how people respond to some treatment (e.g., Augenblick, Niederle, and Sprenger (2015), Rao (2019), DellaVigna, List, Malmendier, and Rao (2022), Hedblom, Hickman, and List (2022), Bodoh-Creed et al. (2023)).

[10] Buchholz, Shum, and Xu (2023) use a similar non-stationary stopping model for taxi-driver labor supply.

Our experiment was *not* designed to test the impact of paying students a certain amount of money to study (e.g., as in Levitt et al. (2016)). Rather, we adopt the student choice model as a basis for an econometric framework under general incentives. Our strategy uses natural field experiments to shape a data-generating process with requisite observables and variation to identify structural parameters governing individual motivation, productivity, and labor-supply costs. Given the alternate methodological focus, our study differs somewhat from common experimental designs. We need not specify a control group as an empirical baseline, but instead structural identification simply requires multiple treatment groups exposed to exogenously differing incentives. Thus, our experiment is similar in spirit to *A/B testing* methods commonly used in marketing and user-experience optimization.

3.1. **Identification Strategy.** Our approach combines standard panel-data methods with recent econometric theory developed for using discrete instruments to quantify continuously varying unobserved heterogeneity (Torgovitsky (2015) and D'Haultfoeuille and Février (2015, 2020)). For intuition on pinning down unobserved student traits, consider a hypothetical "ideal" experiment involving students, $Tabby$ and $Jane$. The researcher obtains two identical clones, call them $Tabby^*$ and $Jane^*$, and during summer break places each of the 4 students into individual observation rooms for a period of two weeks. Inside each room is a desk with a notepad, pencil, and age-appropriate math textbook. There is also a couch with a web-enabled smart-TV and video gaming system for leisure options. Upon entering the observation room, the researcher makes constant piece-rate wage offers—π to $Tabby$ and $Jane$ and $\pi^* > \pi$ to $Tabby^*$ and $Jane^*$—for working through math textbook assignments, with completion defined by some quality criterion. The researcher explains that the children are free to allocate their time in any way they wish, with piece-rate payments to be delivered for the number of exercises successfully completed at the end of two weeks.

Suppose $Tabby$ and $Jane$ complete 5 and 10 math assignments, respectively, while $Tabby^*$ and $Jane^*$ complete 7 and 13. The research team records per-unit study times across completed math assignments for each child, and can infer $\theta_{p,Tabby}$ and $\theta_{p,Jane}$ as panel-data fixed effects. These imply different mean hourly wage rates: say $Tabby$ works fast enough to earn \$15/hour on average, while $Jane$ can garner only \$12/hour on average. All observed differences between $Jane$ and $Jane^*$ are due solely to their piece-rate offers $\pi < \pi^*$, since they are identical and have the same type $(\theta_{p,Jane}, \theta_{m,Jane})$. Since $Jane$ ($Jane^*$) did more work than her same-offer counterpart $Tabby$ ($Tabby^*$) despite lower hourly compensation, $Jane$ must be more willing to allocate time toward math than $Tabby$ (i.e., $\theta_{m,Jane} < \theta_{m,Tabby}$).

The piece-rate shift from π to π^* identifies individual labor-supply elasticities. Moreover, since $\theta_{m,Tabby}$ and $\theta_{m,Jane}$ both interact with a common cost schedule, $c(t)$, differences across the children's choices and labor-supply elasticities can be used to make inference about the shape of $c(t)$, independent of idiosyncratic traits. $Tabby$'s output increased by 40% while $Jane$'s output under the same proportional wage increase rose by only 30%, so marginal costs must be higher from $Jane$'s baseline output of 10 assignments, relative to $Tabby$'s baseline of

5. Inferences about the form of the common cost schedule become richer as the experiment is repeated with a large set of *Tabby*'s classmates, *Clark*, *Anna*, etc. With a complete picture of the shape of the common cost schedule, $c(t)$, the researcher can reverse-engineer each child's motivation type, $\{\theta_{m,Tabby},\ \theta_{m,Jane},\ \theta_{m,Clark},\ \theta_{m,Anna}, \ldots\}$, from the solution of the optimal stopping problem (2), given their observed, optimal choices.

Much is obviously infeasible or unethical about this hypothetical "ideal" experiment. However, one can capture the essential elements with field experimental methods and web-based technologies. Groups of students may be "cloned" through individual-level randomization. While no identical copies of the same child exist, the group-level distributions of observed and unobserved characteristics will be the same. Similarly, rather than sealing students into observation rooms, a web-based learning setup has two considerable advantages. First, a web server can meticulously record time-stamped activities in a non-invasive way that would be impossible otherwise. Second, it allows students to make choices surrounded by the myriad outside options for their time—sports, clubs, music, socializing with friends, chores, etc.—that form a natural part of their regular routine.[11]

3.1.1. *Caveats and Challenges.* Since the researcher cannot observe a student's regular educational activities (e.g., class instruction and graded homework), a question arises: how do we interpret experimentally observed (extracurricular) math activity, given that concurrent, formal coursework and its baseline incentives are unobserved? A major challenge to empirical modelling in many contexts is that the full payoff function encompassing all "carrots" and "sticks" is often difficult to quantify, due to data limitations. Our field experimental design solves this problem by placing many different children on the same footing with known external incentive variation dictated by the researcher. Moreover, structural identification requires only that the distribution of formal coursework activity is uncorrelated with experimental incentives. Thus, individual randomization is crucial to ensure that they are independent of a child's teacher, school, and unobserved external incentives provided by parents/schools/communities. The other central design element is that experimental math activities must be comparable to learning tasks encountered in formal coursework.

Provided the above two criteria are met, concurrent formal coursework and its unobserved external incentives merely changes the interpretation of the motivation parameter somewhat. In the hypothetical, "ideal" experiment, a child's willingness to allocate time toward math activity is judged relative to the baseline of *zero activity*, while in our web-based experiment θ_{mi} represents marginal willingness to allocate *extra time* above and beyond their regular schoolwork. Therefore, structural model estimates remain informative for policy analyses focused on improving academic outcomes *relative to the status quo*. Notwithstanding, note that the interpretation of experimentally inferred productivity, θ_p, hinges only on the similarity between extracurricular incentivized math tasks and formal coursework.

[11]Our web-based research design also provides a proof of concept for powerful new diagnostic tools cheaply available to educators at scale, given recent shifts toward K-12 learning materials being housed online.

Despite this caveat, rich data (discussed in Section 5 and Appendix A) may allow the researcher to move beyond the basic extracurricular interpretation of experimentally inferred θ_m. Recall that $(\theta_{pi}, \theta_{mi})$ represents a fixed effect encompassing all external and internal factors—including default formal coursework commitment—relevant to i's productivity and motivation that are stable over the short-run. Thus, we can project a wealth of student observables (e.g., outside time-use data) on type estimates to study how productivity and motivation are impacted by formal coursework commitments, outside leisure opportunities, demographics, etc. Moreover, structural estimates of unobserved student traits and observed extracurricular math activity may be projected onto exam scores to gain quantitative insights into the "black box" of the learning process, a theme we explore in Section 6.

Two final potential hazards are worthy of note. First, a possible threat to structural identification would arise if students responded to extracurricular incentives by neglecting regular schoolwork. We do not access childrens' academic records due to privacy concerns, but in multiple conversations with our administrator and teacher partners, they universally reported no perceptible reduction in homework completion rates during the sample period. We find strong evidence in our survey data consistent with their reports (see Section 3.3 below). Finally, the thought experiment above glosses over extensive margin choice: what if $Tabby$ spent no time on math under incentive π, while $Tabby^*$ did some math work under π^*? Holding piece-rate incentives fixed, there may be a region of student-type space where either θ_m or θ_p (or both) are prohibitively large to rationalize positive effort. Some group of children may feel that they are too inefficient or too averse to extra work (or both) to respond with positive labor supply. For such students, we cannot point-estimate their 2-dimensional type with a revealed preference approach, but using the whole sample population as a guide, informative bounds can be derived. Our main structural estimator requires only exogenous incentive variation for identification, but our later analyses deal with this challenge via standard Tobit Maximum Likelihood methods (see Sections 5 and 6).

3.2. **Experimental Design Details.** Our field experiment included 1,676 5^{th} and 6^{th} grade students across three demographically distinct school districts in the greater Chicago area. We developed a website with age-appropriate learning tasks professionally designed by experts in mathematics pedagogy. 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.[12] This 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.

3.2.1. *Study Sample.* We partnered with three public school districts in the greater Chicago area for the 2013-2014 academic year. A total of 1,676 $5^{\text{th}}$and 6^{th}grade students participated, with 46% from *District 1*, and 27% each from *District 2* and *District 3*.[13] Although school traits do not directly enter our analyses we summarize them for context in Table OS.3 of the Online Appendix. Relative to the State of Illinois, the state most demographically representative of the U.S. national population at the time of the study, District 1 was above-average on financial resources per student, faculty compensation, teacher qualifications, fraction of budget spent on instruction, and student performance. District 2 was remarkably close to the state averages on most dimensions. District 3 lagged considerably behind the other two on various dimensions, including percent of revenues from local property tax, average teacher qualifications, and student outcomes. District 3 had a relatively high operating budget per-pupil, but this number alone is somewhat misleading. Like many districts serving less affluent communities, it receives additional state funding for factors such as social workers, meal subsidies, and non-instructional support programs.[14] District 3 must also devote significant resources to teaching English as a second language for the 24% of its students who are limited English proficient, in addition to core curriculum subjects.

The local populations these districts serve are similarly ordered by socioeconomic traits. District 1 students are substantially more affluent by income and wealth, with District 2 being closest to state means, and District 3 lagging far behind. The other striking difference is racial sorting of the communities each district serves (see Table 8), which is typical of many US population centers. District 2 has a racially diverse student body, while District 1 serves mostly Whites and Asians, and District 3 serves mostly Blacks and Hispanics.

[12]Experimental procedures underwent stringent ethical vetting by multiple IRBs (at UChicago, UMiami, and BYU). 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 strict data-security measures it would follow. The assent form also allowed parents to opt their child out of the study. 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 also gave each child an opportunity to opt out on their own volition. Language on both assent forms was scrutinized by three research ethics boards. Parents and students recieved multiple notifications—before *and* after data collection—of their right to withdraw from the study. The research team deleted data tied to any child who was opted out of the study.

[13]Our data exclude children in special education, though all who wished were permitted to participate in the incentives program.

[14]In one District 3 school the research team visited, after covering mission critical needs, financial resources to employ full janitorial staff were lacking, so teaching faculty and administrators took turns cleaning the cafeteria room during lunch periods.

3.2.2. *Test Subject Interactions.* We worked closely with 5thand 6thgrade math teachers across the three participating school districts to implement the field experiment. A primary feature of the study was a website on which students could complete up to 80 math learning tasks, each comprised of six practice problems, across five math sub-topics. Students had access to the website for 10 days and could complete as many of the activities as they chose. Our web server monitored students' website use and tallied successful completions.[15] We measured math proficiency using in-class assessments before and after the website was made available. Given our focus on structural identification in this section, we defer discussion on exams and student survey data to Sections 5 and 6, and Appendix A.

3.2.3. *Mathematics Pedagogical Materials.* Proficiency assessments and website content were comprised of professionally developed, age-appropriate math materials. We obtained copies of 46 standardized exams used by various U.S. states over the preceding decade, of which 30 were developed for 5thgraders and 16 were developed for 6thgraders.[16] We split these materials into a bank of 370 unique grade-5 problems and 302 unique grade-6 problems. Finally, we used Common Core Math Standards definitions to categorize each problem into five subject sub-categories: (*i*) *equations and algebraic thinking*, (*ii*) *fractions, proportions, and ratios*, (*iii*) *geometry*, (*iv*) *measurement and probability*, and (*v*) *number system*.[17] We further categorized each math problem by high, medium, and low difficulty, with generous consulting support by pedagogy experts at the UChicago School Math Project.

All 672 problems were pooled to expose 5thand 6thgraders to the same materials. Pooling served multiple purposes. First, it provided a wide swathe of content for studying a diverse student population with considerable pre-existing proficiency heterogeneity. The goal was to achieve a mix of challenging and basic material. Second, it gave us a larger pool of learning materials from which to draw. Third, it facilitated a comparison between age groups, allowing us to cleanly estimate the effect of an additional year of schooling on skill formation.

Of course, this comes at the risk of overwhelming less advanced 5thgraders, and/or failing to sufficiently challenge advanced 6thgraders. Concerns about pooling of students across two age cohorts are mitigated somewhat by the striking similarities in curricula and common-core sample problems across cohorts: grade-6 math curriculum focuses on incremental steps forward from, or applications of, grade-5 curriculum concepts. Online Appendix B.1 and Table OS.1 explain Common Core focus areas by grade, and present a side-by-side harmonization of grade-specific math topics that went into each of our 5 merged sub-categories. Ultimately, the pooling issue boils down to an empirical question: were the offered incentives

[15]Intuitive login credentials were based on the child's first name, last name, grade level, and/or teacher's name. The research team maintained a 24/7 tech-support email to quickly resolve any login problems.

[16]These included *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 Criterion-Referenced Test* (2005).

[17]Common Core subject definitions for 5th/6thgrades (https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf accessed July 2024) differ slightly; our 5-subject classification is a merging of the two (see Appendix B.1).

and pedagogical materials sufficient to attract non-trivial participation from all segments of our sample population? If not, then poor experimental design would be manifest in the form of low statistical power within descriptive analyses and structural estimates. To the contrary, results discussed in Sections 3.3 (esp. Figure 1), 4.3, and 6 (esp. Table 5) demonstrate that there were no undue scientific drawbacks to cohort pooling.

3.2.4. *Website Structure.* Our website was accessible through an individual login credential for each student. The web server automatically tracked and recorded site activity for each child without affecting user experience in any perceivable way. The website provided 80 learning tasks, each consisting of 6 multiple-choice questions from our bank of math problems. Six problems per task were chosen based on feedback from adolescent pilot-study subjects. The passing criterion for completion of each task was at least 5 out of 6 questions answered correctly. Each student was allowed unlimited attempts at a given task, but for each new attempt the ordering of the questions and answer choices were randomly perturbed. Adolescent pilot-study participants universally reported that these measures were enough to make attempts at gaming the system (i.e., repeatedly guessing in rapid succession) unprofitable.

Website learning tasks were organized into 55 general-topic tasks with balanced portfolios of the 5 math topics and 25 topic-specific tasks (5 per topic). Aside from balancing on topical content, questions were selected at random from our bank of math problems, so that relative difficulty was impossible to predict from one activity to the next. After each attempt, an interactive feature provided optional feedback, which the student could choose to skip through or learn from.[18] The web server tracked time by recording a timestamp for each unique page view. Since only one math problem appears per page view within each learning task, we get a high-frequency log of work times for each child.[19] The website logged successful completions into a database, and reported completed tasks, piece-rate incentives, and current earnings to the user. Thus, we can obtain data on total website time T_i, task accomplishment A_i, and a rich panel of within-child, task-specific, work times $\{\tau_{a_i}\}_{a_i=1}^{A_i}$.

The website was designed to be mobile-device friendly to accommodate children with various types of internet connections at home. Roughly 3/4 of pageload requests originated from computers (including laptops), about 1/5 were from tablets, and about 1/20 of pageloads were from smartphones. One potential concern is that limited internet access may have unduly influenced our results by inhibiting participation for some students. Although 7% of students reported not having a regular internet connection at home, they were not statistically less likely to participate on the website than others, suggesting they had other ways of

[18]There was also an instructive component built from math textbook glossaries (by the University of Chicago School Math Project, `ucsmp.uchicago.edu`) and interactive example problems. This instructive component was clearly marked as non-paid, but it provided an option for students to invest in their income generation capability. Less than 2% of overall page-view time was logged on the instructive portion of the website.

[19]One technical concern was a small number of spurious page-view times that result when a child closed her web browser in the middle of a task without logging off. We replace these with student-sub-category-problem mean work times, using a procedure proposed by (C. Cotton et al., 2022, Online Appendix).

accessing the internet.[20] Having a regular internet connection at home was not a statistically significant predictor of website task completion, after controlling for school district, socioeconomic status, regular homework time, math attitude, and incentives (see Online Appendix B.2). These results strongly suggest that internet access was not a major concern, due to our mobile-friendly website design, the network of 11 public libraries serving our sample population, and various other available options (extended family, school library, etc.).

3.2.5. *Incentives and Randomization.* We offered linear incentives (constant piece-rate) simple enough for adolescents to easily understand: $\Pi_i(a_i) = (\pi_0^i + \pi_1^i a_i)\mathbb{1}(a_i \geq 2)$. To ensure a within-student panel of data, we informed students that they must complete at least 2 website 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.[21] Moving forward, we use superscripts to denote student i's contract assignment, while subscripts denote the fixed payment parameters for each contract; in other words, the statement $(\pi_0^i, \pi_1^i) = (\pi_{0j}, \pi_{1j})$ indicates that student i operated under contract-j incentives. Our website achieved incentive salience by prominently advertising piece-rates to users on the home page, as well as total accrued earnings (with real-time updates), and remaining potential earnings. Adequate subject motivation depends on the ratio of payoffs to work for each learning task. Partitioning them into small units (6 problems) encouraged participation and facilitated precise panel-data inference on θ_{pi}. Note that our three offers vary significantly in proportional terms: contracts 2 and 3 had piece-rate raises of 33% and 66.7%, respectively, over contract 1.

Our block randomization separated students into race-gender-school-grade bins. Within each, we ordered students 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. Table OS.4 (Online Appendix) shows that our treatment assignment was independent of all balancing variables. 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.[22] Letters also promised prompt delivery of payments within 2 weeks following the end of the experiment.

[20] The 95% confidence interval for participation rate among students without a regular home internet connection, $[0.318, 0.495]$, contains the participation rate for the overall sample of 0.447. Among this group, computer-based pageloads were 14% lower and smartphone pageloads were higher by a similar margin.

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

[22] A potential concern is whether students shared their login credentials with others. While impossible for us to fully verify, various factors suggest not. First, $>95\%$ students in a grade cohort received login credentials, so they would not likely have been willing to do work for someone else if they could do the same work for themselves for pay. Second, roughly $1/2$ of sampled students declined any math work on the website, ruling out widespread login sharing which would have inflated work volumes recorded by the web server. Third, we see a strong and statistically significant relationship between completed website learning tasks and gains in math proficiency (see Tables 8 and 7). This relationship persists after controlling for a wealth of student observables, and suggests that completed website tasks reflect their own work, and not someone else's.

The design of our website and incentives had several advantages. First, we incentivized successful completion of learning tasks rather than time spent on these tasks. This is consistent with typical academic settings where students are rewarded or punished (by schools/parents) based on whether they complete homework assignments. Our focus on short-run, at-home math practice (analogous to a short homework assignment) rather than long-term outcomes like semester-end grades, made student decisions in our sample consistent with their day-to-day choices on homework completion. Second, our small window of effort—in terms of size of incentivized tasks and payment timeline—minimized the temporal gap between effort and reward in order to maximize salience. Recent research (e.g., Bettinger (2012) and Levitt et al. (2016)) has shown that incentives are more effective when rewards closely follow actions. Third, as in many contemporary web-based homework platforms used by professional K-12 educators, we allow students multiple attempts at passing each learning task. This is consistent with our model of stochastic completion times: a rational student clicks "submit" on an attempt only if she believes she may pass, but she doesn't know how much additional time it will take until the website reports that her last attempt was successful.

3.2.6. *Experiment Timeline.* The experiment proceeded 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.

3.2.7. *Classroom Tests and Surveys.* In addition to website activity logs, we collected student test-score and survey data. These are not needed for structural identification, but they enable secondary analyses in Sections 5 and 6 where we enhance interpretability of structural parameters by decomposing student type estimates, (θ_p, θ_m), and investigate production of durable skills. Teachers in participating classrooms proctored standardized math assessments based on common-core-classified sample problems as pre- and post-tests for the experiment. Exam content was drawn from the same sources as website materials, as described in Section 3.2.3 above. Each exam contained 36 questions, chosen to balance 5^{th}- and 6^{th}-grade content, sub-topics, and difficulty level. All students were given 35 minutes to complete as many of the 36 problems as possible. Students also completed surveys, which collected information on a myriad of individual factors, including attitudes, extracurricular activities, regular 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 our participants resided. Full details on in-class assessments and surveys, including descriptive statistics, are provided in Section 5, Section 6, and Appendix A.1.

3.3. **Descriptive Analyses of Website Activity.** Table 1 displays descriptive statistics of math website activity. It will be useful to 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. 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. Overall, we observe 749 active students who completed 16,740 learning tasks (i.e., 84,000–100,000 math problems correctly solved) across roughly 30,000 attempts and 2,000 child-hours during our 10-day sample period.

To place these numbers into perspective, first recall that website activity was extracurricular, being separate from a child's regular schoolwork regimen. For a basis of comparison, we compiled survey data on school homework time per day (across all school subjects): on the pre-survey we asked students about their homework time during a "typical week", and on the post-survey we repeated the same question but referring specifically to the sample period. One possible threat to identification would be if students responded to the extracurricular financial incentives by neglecting schoolwork in proportion to their website activity. In conversations with participating administrators and teachers, they universally reported a firm impression that students did not reduce the amount of turned-in homework assignments during the sample period. Our survey data corroborate this claim: among active students, mean homework time reports across the pre-survey and post-survey differed only by a small margin (3.97%), and the difference was statistically insignificant (p-value$=0.156$). Table 1 reports homework time numbers averaged across pre- and post-survey reports.

Aside from acting as a robustness check, this result contextualizes the magnitude of observed website activity. Assuming mathematics accounted for 25%–50% of daily homework time implies the average (median) website math time per day among active students would have represented an increase of 37%–74% (24%–48%) in daily math activity.[23] Active students reported 22.1% more daily homework time than marginals/inactives, and a two-sample t-test rejects the null hypothesis of active vs. marginal/inactive mean equality (p-value 3.5×10^{-38}). Other indicators in our data also point to a strong positive relationship between daily homework times and willingness to engage in extracurricular math. We find positive Spearman rank correlations between daily homework time and three different measures of website activity: (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}). Finally, students were asked on our surveys to rate how often they miss homework assignment deadlines at school; their responses have a statistically significant negative relationship with choices

[23]For an alternate benchmark, we discussed our findings with a math education consultant employed by a Midwestern US state. The consultant opined that 72 extra math problems within a 10 days (the active student median) would be an increase of 50%–100% in homework volume for a typical $5^{\text{th}}/6^{\text{th}}$grade student.

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 $\mathbb{1}(A_i \geq 2)$	0.447	0	0.497	1,676	0.422	0.453	0.466
Marginal Students $\mathbb{1}(A_i = 1)$	0.056	0	0.230	1,676	0.072	0.043	0.054
Inactive Students $\mathbb{1}(A_i = 0)$	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	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.)	157.05	102.85	152.45	749	122.74	160.13	184.96
Within-Child Avg. Time Per Comp. Task (min.)	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)*						

of time spent on our math website, with a Spearman rank correlation of -0.265 (p-value 2.6×10^{-26}). These results suggest a meaningful link between our website data and unobservable differences across students that drive disparate choices and outcomes over time.

FIGURE 1. Website Choices and Performance

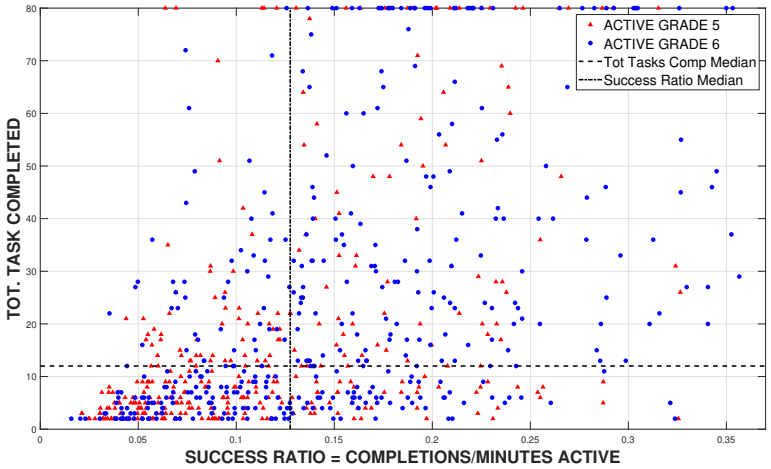


Figure 1 provides preliminary insights into unobserved student heterogeneity, based on field-experimental observables. The horizontal axis is number of website learning tasks completed, A_i, and can be thought of as analogous to measures typically available in observational data (e.g., assignment completions, GPA, exam scores). The horizontal axis is *success ratios*, or task completions per unit of learning time, A_i/T_i, and is typically not available in observational education data. Both measures illustrate vast heterogeneity, and while 6^{th}graders (circles) are more efficient than 5^{th}graders (triangles), on average, both groups have significant representation across a common support, providing further assurance that our pooling the two age cohorts within the experiment was reasonable.

The scatter-plot provides intriguing reduced-form evidence on unobserved study productivity and motivation. First, the northwest quadrant—low success ratio but high work volume—and the southeast quadrant—high success ratio but low work volume—are both well populated. While the rank correlation between success ratio and task completion is (unsurprisingly) high at 0.551 (p-value 1.4×10^{-60}), the rank correlation between success ratio and website time choice is surprisingly low, at 0.067 (p-value 0.067). If we focus on students who completed at or above the median output A_i plus one standard deviation, that gives us a cutoff of 36 completed tasks on the vertical axis. Among this high-achiever group, which is often labelled as "gifted" or "talented" based on traditional observables, a striking feature of the plot is vast conditional productivity heterogeneity. This reduced-form finding from our field experiment provides a new window into the "black box" of academic success.

A key prediction of our model is that sufficient strength on either trait, θ_p or θ_m, is enough to drive high observed achievement. A very inefficient child (i.e., high θ_p) may still produce high work volume with sufficiently high motivation (i.e., low θ_m), and vice versa. Figure 1 is strongly consistent with this idea, and two 6^{th}-grade data-points in the scatter-plot, call them *child 1* $(0.057, 36)$ and *child 2* $(0.353, 37)$, vividly illustrate. Conventional wisdom would label *child 2* as "slightly more motivated" for having exceeded *child 1*'s apparent effort level by a margin of one more completed learning task. Our model and field-experimental data paint a very different picture. Despite requiring over six-fold more time to achieve each incentivized success, *child 1* nearly matched *child 2*'s output, and is therefore *vastly more motivated*. Moreover, this begs the question of how different *child 1*'s life might be under some intervention that narrowed the productivity gap between him/her and *child 2*.

4. Identification and Estimation of the Student Choice Model

The primary structural primitives of the study-leisure model are idiosyncratic productivity (θ_{pi}) and motivation (θ_{mi}) parameters, and the cost function $c(t)$. Additional structural parameters include τ_0, τ_1, φ, and work-time shock CDFs $F_u(u_{a_i}|\theta_{pi})$. We now discuss identification and sketch out a two-stage GMM estimator to implement our empirical strategy.

4.1. Stage-1 Estimation: Productivity and Work-Time Shocks.

Our approach follows standard methods using the within-child panel structure of work-time data, $\{\tau_{a_i}\}_{a_i=1}^{A_i}$.

Log transforming equation (1) produces a linear-in-parameters regression equation, $\log(\tau_{a_i}) = \log(\tau_0) + \log(\tau_1)\mathbb{1}(a_i = 1) + \log(\theta_{pi}) - \varphi\log(a_i) + \log(u_{a_i}),\ a_i = 1, \ldots, A_i,\ \{i|A_i \geq 1\}$, where θ_{pi} enters as a student fixed-effect, and $(\tau_0, \tau_1, \varphi)$ enter as intercept and slope terms. We estimate these parameters and fixed effects through a standard differencing approach.

For estimation of heteroskedastic study-time shock CDFs $F_u(u|\theta_p)$, we first compute fitted residuals, $\hat{u}_{a_i} = \tau_{a_i}/(\hat{\tau}_0\hat{\tau}_1^{\mathbb{1}(a_i=1)}\hat{\theta}_{pi}a_i^{-\hat{\varphi}}),\ a_i = 1, \ldots, A_i,\ \{i|A_i \geq 1\}$, and we partition the support of $\hat{\theta}_{pi}$ into 5 sub-intervals of equal length, $I_j^p,\ j = 1, \ldots, 5$.[24] Then, we split fitted residuals into 5 sub-samples $\{\{\hat{u}_{a_i}\}_{a_i=1}^{A_i}\}_{\{i|\hat{\theta}_{pi} \in I_j^p\}},\ j = 1, \ldots, 5$, and we smooth the corresponding empirical CDFs using a flexible cubic B-spline form $\hat{F}_u(u|I_j^p; \boldsymbol{\gamma}_{uj}) = \sum_{k=1}^{7} \gamma_{ujk}\mathcal{B}_{ujk}(u),\ j = 1, \ldots, 5$.[25] Estimates are consistent with heteroskedastic shocks: students who take longer to solve math problems also have larger work-time variances than their more efficient counterparts.

Stage-1 model components can be separately pre-estimated under the assumption,

Assumption 4. Study-time shocks U_{a_i} are conditionally independent of motivation, Θ_{mi}, given child i's productivity type θ_{pi}.

Intuitively, this means that a child's motivation parameter θ_{mi} operates only on her decision to devote time to math or the outside option. Conditional on allocating time to math, she invests full cognitive resources into the incentivized task and operates at her production possibility frontier, modulo random, unpredictable shocks. Under this assumption, stage-1 parameters including $\{\hat{\theta}_{pi}\}_{\{i|\hat{\theta}_{pi} \in I_j^p\}}$, $\hat{\tau}_0$, $\hat{\tau}_1$, $\hat{\varphi}$, and $\hat{F}_u(u|I_j^p; \hat{\boldsymbol{\gamma}}_{uj})$, $j = 1, \ldots, 5$, can be treated as known (and fixed) during stage-2 estimation. This provides needed tractability by drastically reducing parameter-space dimension and computational burden.

One challenge is that individual fixed-effect estimates have differing variances due to the unbalanced panel: A_i varies across active students, and higher values lead to more precisely measured $\hat{\theta}_{pi}$.[26] In our secondary analyses in Section 5 we use inverse-variance weighting, and in Section 6 we implement Feasible Generalized Least Squares methods and robust standard errors to address any heteroskedasticity issues that arise from unbalanced panel estimation in Stages 1 and 2. A final challenge is that student fixed effects can only be point-identified for active students. This problem plays only a minor role in our stage-2 structural estimator, based on exogenous incentive variation, and is straightforward to deal with in our secondary analyses in Section 5 by use of a standard Tobit Maximum Likelihood approach.

4.2. **Stage-2 Estimation: Labor-Supply.** Formal identification of idiosyncratic student labor-supply elasticities builds on ideas developed by Torgovitsky (2015) and D'Haultfoeuille and Février (2015, 2020). These papers explore conditions under which discrete instruments

[24] 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.

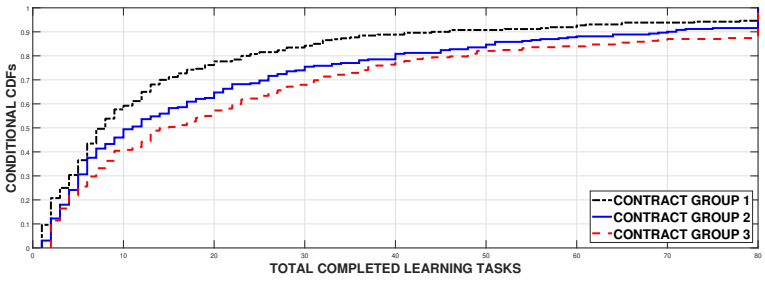
[25] Basis functions $\mathcal{B}_{ujk}$ are determined by the Cox-de Boor formula and a pre-specified knot vector spanning $supp(F_u)$. We chose 4 knots, uniformly spaced in quantile ranks. After constraining the endpoints this left 5 free parameters, which achieved a remarkably tight model fit depicted in Figure OS.1 (online appendix).

[26] Cross-student variation in panel length is central to identification in Stage 2, and the unbalanced panel problem exists independently of whether stage-1 objects are pre-estimated or not.

are sufficient to nonparametrically identify a continuum of unobserved heterogeneity (θ_{mi} in our case) without *a priori* functional form restrictions on the cost function $c(t)$. Stage-2 identification relies on exogenous variation in observable choices, which our experimental design achieves via randomized incentives (π_{0j}, π_{1j}) across groups of adolescents, $j = 1, 2, 3$, who otherwise have identical distributions of unobserved traits. Table 1 and Figure 2 show descriptive evidence of the exogenous variation on which our identification strategy is based. The final three columns in the table depict a steady increase of activity level, learning task completion, and time spent on the website between contract groups 1, 2, and 3.

There are two basic tasks the Stage-2 structural estimator must accomplish: (i) pin down idiosyncratic labor-supply elasticities determined by θ_{mi}, and (ii) trace out the curvature of the labor-supply cost function $c(t)$. Conditional CDFs of A_i, plotted in Figure 2, are primary data moments relevant to these two tasks. Two artifacts of the figure are especially illustrative for identification. First, the three CDFs follow a stochastic dominance ordering that the model predicts, given the progression of our piece-rate incentives across contracts. A formal nonparametric stochastic dominance test proposed by Barrett and Donald (2003) reveals that the null hypotheses of pairwise equality among the three CDFs are rejected in favor of first-order dominance. For accomplishing task (i), randomization combined with monotonicity in the mapping between θ_{mi} and A_i (holding θ_{pi} fixed) implies that individual labor-supply responses to incentive shifts can be pinned down by quantile differences across the three conditional CDFs. A child at the median output level under contract 1 would, on average, attain the median output level under contracts 2 or 3 as well, since the three contract groups have the same underlying distribution of unobserved types (Θ_m, Θ_p).

FIGURE 2. Math Website Output by Contract Group



Notes: Null hypotheses of pairwise distributional equality are rejected by the Barrett-Donald (2003) test (using 100,000 bootstrap samples) in favor of the alternate hypothesis of first-order dominance with the following p-values: *Group 1* vs *Group 2* p-value=0.002; *Group 1* vs *Group 3* p-value$<10^{-5}$; *Group 2* vs *Group 3* p-value=0.064.

Second, Figure 2 depicts two unequal stochastic shifts, despite the marginal wage difference (\$0.25 per completed task) between contracts 1 and 2 being the same as the difference between contracts 2 and 3. This fact helps with task (ii): cost curvature (i.e., $c''(t) > 0, \ \forall t$) implies a model prediction that quantile differences in work volume when shifting from contract 1 to contract 2 should be larger than when shifting from contract 2 to contract 3. Raw data confirm this prediction: the first integrated quantile difference, $\int_0^1 \left[\hat{G}_a^{-1}(r|\pi_{12}) - \hat{G}_a^{-1}(r|\pi_{11})\right] dr$,

is 5.77 additional learning tasks, on average, while the difference for the other two contracts, $\int_0^1 \left[\hat{G}_a^{-1}(r|\pi_{13}) - \hat{G}_a^{-1}(r|\pi_{12})\right] dr$, implies an average of 3.82 additional learning tasks, or a 33% reduction in labor-supply response. This double difference helps pin down cost curvature.[27]

4.2.1. *Simulated GMM Estimator Overview.* Our estimator is built on functional representations of these counterfactual quantile comparisons. We use a flexible cubic B-Spline specification of costs, $\hat{c}(t;\boldsymbol{\gamma}_c) = \sum_{k=1}^{k_C+3} \gamma_{ck}\mathcal{B}_{ck}(t)$, with knot vector $\boldsymbol{\kappa}_c = \{\kappa_{c1}, \kappa_{c2}, \ldots, \kappa_{c,K+1}\}$.[28] For any fixed shape of the cost function (conforming to Assumption 2), the researcher can employ techniques in the spirit of Hotz and Miller (1993) and Guerre, Perrigne, and Vuong (2000) to reverse-engineer a child's motivation type θ_{mi} from her observable choices (T_i, A_i), using equations (2) and (4). Consider child i, whose learning-task volume A_i was at quantile rank r_i in Contract Group 1. We can repeatedly simulate sequences of work-time shocks from $F_u(u|\theta_{pi})$, and associated work times using (known) θ_{pi}, τ_0, τ_1, and φ. Then, holding fixed the cost-function parameters $\boldsymbol{\gamma}_c$ and child i's actual incentives, denoted (π_0^i, π_1^i), we find θ_{mi} such that the distribution of her optimal choices imply mean stopping time equal to i's observed T_i.[29] When solving for optimal stopping choices, we employ a new method recently proposed by Hamilton, Hickman, and Mohie (2024) for tractable computation of non-stationary dynamic programming problems with history dependence.

Observed choices are also informative of cost curvature. First, A_i contributes to the empirical CDF of work volume under i's actual contract-1 assignment, $\hat{G}_a(a|\pi_{11}) = \sum_{i=1}^N \frac{\mathbb{1}\left[A_i \leq a \;\&\; \pi_1^i = \pi_{11}\right]}{\sum_{i=1}^N \mathbb{1}\left[\pi_1^i = \pi_{11}\right]}$. Second, we can also simulate a sequence of *counterfactual work-volume choices*, $\{\tilde{A}_{i2s}\}_{s=1}^S$ under contract 2, and $\{\tilde{A}_{i3s}\}_{s=1}^S$ under contract 3. These depend on $F_u(u|\theta_{pi})$, θ_{pi}, τ_0, τ_1, and φ (known and fixed), and on the shape of the cost function $\hat{c}(\cdot;\boldsymbol{\gamma}_c)$ through equations (1)–(4). Simulated counterfactuals pin down model-generated CDFs of work volume under assignment to contracts 2 and 3 through the following: $\tilde{G}_a(a|\pi_{1j};\boldsymbol{\gamma}_c) = \sum_{i=1}^N \sum_{s=1}^S \frac{\mathbb{1}\left[\tilde{A}_{ijs} \leq a \;\&\; \pi_1^i \neq \pi_{1j}\right]}{\sum_{i=1}^N \mathbb{1}\left[\pi_1^i \neq \pi_{1j}\right] \times S}$, $j = 2, 3$. Thus, i's observed choices (T_i, A_i) contribute to the empirical CDF of her actual group, $\hat{G}_a(a|\pi_{11})$, and they also contribute to the model-generated CDFs $\tilde{G}_a(a|\pi_{12};\boldsymbol{\gamma}_c)$ and $\tilde{G}_a(a|\pi_{13};\boldsymbol{\gamma}_c)$. Of course, there is nothing special about i being in Contract Group 1, and similar logic can be applied to all active students.

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 empirical CDFs $\hat{G}_a(\cdot|\pi_{1j})$, $j = 1, 2, 3$,

[27] Torgovitsky (2015) and D'Haultfoeuille and Février (2015, 2020) show that rich utility curvature information is encoded within the curvature of a single quantile difference, $G_a^{-1}(r|\pi_{1j}) - G_a^{-1}(r|\pi_{1j'})$, as well.

[28] We chose a knot vector with $K_c = 7$ sub-intervals, or 8 knots spaced uniformly in quantile-ranks of T in order to evenly spread the influence of data over the various parameters $\boldsymbol{\gamma}_c$. After imposing Assumption 2, this left 8 free parameters characterizing labor-supply costs. While B-splines may produce a semi-nonparametric sieve estimator—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 instead view our 10-parameter B-spline cost function as a flexible but fixed parametric form.

[29] In a slight shift in notation, here we use T_i to denote i's total work time through all *completed learning tasks*, net of any time spent on unfinished work tasks. While this choice leaves a small amount of empirical information on the table, it lends a great deal of computational tractability to the problem by drastically reducing the number of continuation value function evaluations during when simulating the model.

and their model-generated counterparts, $\tilde{G}_a(\cdot|\pi_{1j};\boldsymbol{\gamma}_c),\ j=1,2,3$:

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_c = \text{argmin}\Bigg\{ & \textstyle\sum_{l=1}^{L}\sum_{j=1}^{3}\Big(\hat{G}_a(a_l|\pi_{1j}) - \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\Big)^2 \\ & +\omega_0\times\Big(\hat{\overline{G}}_a^{90}(a_l|\pi_{1j}) - \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\Big)^2\times\mathbb{1}\Big[\hat{\overline{G}}_a^{90}(a_l|\pi_{1j}) < \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\Big] \\ & +\omega_0\times\Big(\hat{\underline{G}}_a^{90}(a_l|\pi_{1j}) - \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\Big)^2\times\mathbb{1}\Big[\hat{\underline{G}}_a^{90}(a_l|\pi_{1j}) > \tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\Big]\Bigg\}\end{aligned} \tag{5}$$

$$s.t.\ \gamma_{c1}=0;\ \gamma_{c2}=(\kappa_{c2}-\kappa_{c1})/3;\ \gamma_{ck}-\gamma_{c,k-1}>0,\ k=2,\ldots,K_c+3,;\ \tfrac{\gamma_{c,k+1}-\gamma_{ck}}{\kappa_{c,k+1}-\kappa_{ck}}-\tfrac{\gamma_{ck}-\gamma_{c,k-1}}{\kappa_{ck}-\kappa_{c,k-1}}>0\ k=2,\ldots K_c+2.$$

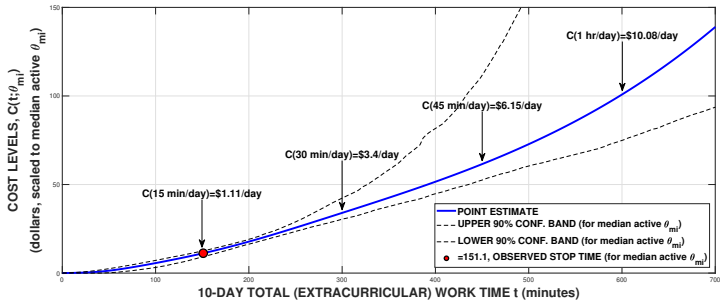
The first line of the objective is the primary least squares moment conditions, and the last two lines are "guardrail" moment conditions for numerical stability. $\hat{\overline{G}}_a^{90}(a_l|\pi_{1j})$ and $\hat{\underline{G}}_a^{90}(a_l|\pi_{1j})$ are the (interpolated) point-wise 90% confidence bounds of the empirical CDFs, and ω_0 is a penalty parameter. Guardrail conditions help the solver to avoid becoming stuck at local optima in $\boldsymbol{\gamma}_c$-space by imposing a quadratic penalty in regions where the model-generated CDFs $\tilde{G}_a$ differ from the empirical analogs by more than the 90% confidence bounds. Otherwise, they play no role. Constraints enforce boundary value ($c(0)=0$), boundary derivative ($c'(0)=1$), monotonicity ($c'(t)>0$), and convexity ($c''(t)>0$), respectively.

4.2.2. *Correcting for Sample Selection.* There are two minor complications regarding mass points at the extremes of the sample. First, a small mass of students achieve full output $A_i=80$ on the website. This issue is straightforward to deal with: since their productivity types θ_p are known, we compute multiple hypothetical θ_m values for each using values of A drawn from an extrapolated upper tail of the distribution $G_a(a|\pi_{1j})$. Details are discussed in Appendix B.3.1 (see also Figure OS.2).

The other minor challenge relates to non-active students who completed fewer than 2 learning tasks $A_i\leq 1$ (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 ($A_i\geq 2$) 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, $\tilde{G}_a(a|\pi_{1j})$, $a\geq 2$, are based on underlying sets of students with comparable unobserved types.

4.3. **Structural Estimates.** Structural parameter estimates and bootstrapped confidence intervals are in Table 10 in Appendix A.3. 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. The most interpretable structural primitive is $\hat{\varphi}=0.0788$, the

FIGURE 3. Time Supply Cost & Marginal Cost Estimates



experience-curve parameter. This estimate is statistically significant, but implies only minor short-term productivity gains: for a baseline of current work a_i, mean per-unit completion time on the $(2a_i)^{\text{th}}$ task (i.e., doubling volume) drops by only 5.32%.

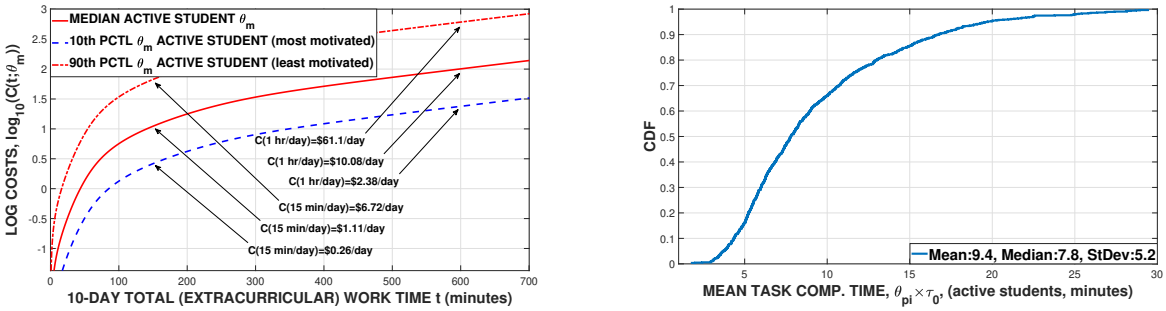
Figure OS.1 in the online supplement presents estimated densities of heteroskedastic production time shocks. We find that productivity shock distributions for less productive (i.e., higher θ_p) students are roughly mean-preserving spreads of the corresponding shocks for their more productive (lower θ_p) 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 between 44% below and 89% above median completion time.

4.3.1. *Cost Schedule.* Figure 3 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 sparse. Figure OS.3 (online appendix) depicts goodness of fit of our flexible structural model by comparing empirical CDFs and model-predicted CDFs of work volume by contract group. Overall, the structural model does remarkably well at matching patterns in the data, especially for contract group 2 where the richest set of counterfactual comparisons are available (i.e., with *both* higher and lower incentives).

We estimate a high degree of curvature in the cost function $c\,(t; \widehat{\boldsymbol{\gamma}}_c)$. Figure 3 labels cost levels at regular intervals to illustrate this point. The child whose cost schedule is depicted chose a total time commitment to our offered extracurricular math activities of 151.1 minutes, or just over 15 extra minutes per day over our sample period. At this level of sustained additional math activity, this median child would have incurred a daily utility cost of just over \$1.11 per day. A doubling of this marginal math-time allocation roughly triples costs, and an increase up to 1 hour/day raises utility costs by an order of magnitude.

These numbers mask subtle agency issues within educational contexts. Labeled cost levels are monetary transfers that exactly offset costs associated with a certain commitment to marginal math activity beyond status-quo schoolwork. A principal who can force this median student into an hour of extra math study per day to reach competency standards, could make the child whole again (i.e., zero surplus) with a daily transfer of \$10.08. However, this

FIGURE 4. Motivation/Productivity Heterogeneity



hypothetical assumes access to the child's private information and a means of compelling him/her to some level of effort increase. Otherwise, the principal must offer incentives and allow the child to optimize, in which case he/she will choose an optimal stopping time that ensures positive surplus. The depicted child in Figure 3 was in contract group 3, and completed 28 learning tasks in 15.1 minutes/day, resulting in a surplus of $28. In that sense, cost levels depicted are actually deceptively low: for the median student to rationally choose an hour of extra daily math under private information and limited commitment, the principal would have to offer incentives far exceeding $10.08 daily.

4.3.2. *Motivation and Productivity Heterogeneity.* The left panel of Figure 4 illustrates cost variation *across students*. The figure depicts cost schedules scaled to θ_m types at the 10^{th} percentile (i.e., highly motivated), median, and 90^{th} percentile (i.e., less motivated) of active students, where we have log-transformed costs to facilitate a graphical comparison. We find dramatic 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 striking variation only adds to the challenges of the information- and commitment-constrained principal described above: not knowing who is highly motivated and who is not, 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. Alternatively, 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. Of course, under piece-rate academic incentives, motivation is only one piece of the puzzle of student choice. The right panel of Figure 4 depicts productivity differences: the 10-90 range entails more than a 4-fold change in mean task completion time, and in turn, an equivalent 4-fold range of hourly compensation, holding piece-rate incentives fixed.

5. EXTERNAL INFLUENCES ON MOTIVATION AND PRODUCTIVITY

In this section and the next, we use a series of secondary analyses to show how structural type estimates $(\hat{\theta}_{pi}, \hat{\theta}_{mi})$ open a new window into what differentiates individuals' learning processes. Motivation heterogeneity may be driven by internal psychic costs of working on math, or by external factors (e.g., quality and variety of leisure activities) which shape

opportunity costs of foregone time. Similarly, 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). This begs the question of, how much variation in productivity/motivation is explainable by external factors, and is there a role for education policy to influence the type of learner a child becomes?

We model θ_p and θ_m as comprising internal and external components as follows:

$$\log(\theta_{pi}) = \boldsymbol{X}_{pi}\boldsymbol{\beta}_p + \eta_{pi}, \quad \text{and} \quad \log(\theta_{mi}) = \boldsymbol{X}_{mi}\boldsymbol{\beta}_m + \eta_{mi}. \tag{6}$$

Here, $\boldsymbol{X}_{pi}$ and $\boldsymbol{X}_{mi}$ are vectors of student-level covariates, while (η_{pi}, η_{mi}) represent the truly idiosyncratic component of student i's unobserved traits $(\theta_{pi}, \theta_{mi})$. One obstacle to overcome is truncation of the outcome variable: while data on $(\boldsymbol{X}_{pi}, \boldsymbol{X}_{mi})$ is available for all i, the left-hand variables $(\log(\theta_{pi}), \log(\theta_{mi}))$ are known precisely only for active students $(A_i \geq 2)$. For marginal/inactive students $(A_i < 2)$, structural estimates from Section 4.3 allow us to bound their (θ_p, θ_m) types using known contract-specific selection thresholds, $\underline{\Theta}_m(\theta_p; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c)$ and $\underline{\Theta}_p(\theta_m; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c)$, where $\underline{\Theta}_p(\theta_m; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c) = \underline{\Theta}_m^{-1}(\theta_m; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c)$, $j = 1, 2, 3$.[30] Intuitively, $\underline{\Theta}_m(\theta_p; \pi_{0j}, \pi_{1j}, \widehat{\boldsymbol{\gamma}}_c)$ is the marginal motivation type willing to complete at least 2 website tasks, given productivity θ_p and incentives from contract j. The thresholds imply

$$\begin{aligned} \log(\theta_{pi}) &\geq \log\left(\underline{\Theta}_p(\boldsymbol{X}_{pi}\boldsymbol{\beta}_p + \eta_{pi}; \pi_0^i, \pi_1^i, \widehat{\boldsymbol{\gamma}}_c)\right), \quad \text{and} \\ \log(\theta_{mi}) &\geq \log\left(\underline{\Theta}_m(\boldsymbol{X}_{mi}\boldsymbol{\beta}_m + \eta_{mi}; \pi_0^i, \pi_1^i, \widehat{\boldsymbol{\gamma}}_c)\right) \end{aligned} \tag{7}$$

for marginal/inactive students. We solve the truncated dependent variable issue by a standard approach: adopting a parametric assumption on the joint distribution of errors.

Assumption 5. Residual productivity and motivation are normal, $(\eta_{pi}, \eta_{mi}) \sim BVN(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, with variance-covariance structure being a function of race and gender: $\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{pi}^2 & \sigma_{pmi} \\ \sigma_{pmi} & \sigma_{mi}^2 \end{bmatrix}$, $\sigma_{ji} = \sigma_{j0} + \sigma_{j1} fem_i + \sigma_{j2} black_i + \sigma_{j3} hispanic_i, \quad j = p, m, pm.$

Assumption 5 with equations (6) and (7) facilitate a bivariate Tobit Maximum Likelihood estimator, defined as the arg max of the following log-likelihood function:

$$\begin{aligned} \left[\widehat{\boldsymbol{\beta}}_p, \widehat{\boldsymbol{\beta}}_m, \widehat{\boldsymbol{\Sigma}}\right] = \text{argmax} \Bigg\{ & \textstyle\sum_{i=1}^N \mathbb{1}(A_i \geq 2)\omega_{di} \log\left(f_{\eta_p, \eta_m}(\boldsymbol{X}_{pi}\boldsymbol{\beta}_p, \boldsymbol{X}_{pi}\boldsymbol{\beta}_p; \boldsymbol{\Sigma}_i)\right) \\ & + \mathbb{1}(A_i < 2)\omega_{di} \log\left(\Pr\left[\log(\theta_m) > \log\left[\underline{\Theta}_m(\theta_p; b_i, \pi_{1i}, \widehat{\boldsymbol{\gamma}}_c)\right] \middle| \boldsymbol{X}_{pi}, \boldsymbol{X}_{mi}; \boldsymbol{\beta}_p, \boldsymbol{\beta}_m, \boldsymbol{\Sigma}_i\right]\right) \Bigg\}, \end{aligned} \tag{8}$$

where the ω's are 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 $A_i < 2$. For tractability, we compute the selection probability above by simulation.[31]

[30] Thresholds can be estimated as the northeast boundary of the convex hull of $(\hat{\theta}_{pi}, \hat{\theta}_{mi})$, given $A_i \geq 2$.

[31] More concretely, we begin by simulating 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 $(\tilde{\boldsymbol{\theta}}_{pki}, \tilde{\boldsymbol{\theta}}_{mki})$ (see equation (6)), and we use it to compute selection probability in the second line of (8). A simple way of doing so would be a sample mean of indicator

In order to attach a causal interpretation to estimates of the parameters in equation (6), we require the following additional assumption on the error structure:

Assumption 6. $E[\boldsymbol{X}_{ji}^{\top}\eta_{ji}] = \mathbf{0}$, $j = p, m$.

After presenting results we return to a discussion on the plausibility of this assumption in Section 5.1.7 below. In our main specification, $\boldsymbol{X}_{pi}$ contains an intercept and the following: indicators for *gender*, *race*, *grade* (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 proxies for i's neighborhood of residence: *mean household income* and *fraction minors with no health insurance*. These last two variables are measured at the neighborhood (Census Block Group) level, of which there are 161 across our 3 school districts. They serve as proxies 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 continuous variables *mobilefrac* and *tabletfrac*, being fractions of website page-loads from a smartphone or tablet device (desktop computer is the omitted category). They give us a window into how kids cope with learning activity under resource constraints. Finally, as a robustness check we will add four additional controls (described below) from a parent survey.

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. Although our focus in this section is on external forces acting on child learning, the first four additional motivation controls measure internal attitudes toward math specifically—indicators for whether the child reported math as their *favorite* academic subject or *least favorite*—and attitudes toward effort or work in general—two survey-elicited indices for *extrinsic motivation* and *intrinsic motivation*. 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 supervised leisure time* or how much of a child's leisure is spent 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*.

functions, $\Pr\left[\theta_m > \underline{\Theta}_m(\theta_p; b_i, \pi_{1i}, \widehat{\gamma}_c)\middle|\boldsymbol{X}_{pi}, \boldsymbol{X}_{mi}; \boldsymbol{\beta}_p, \boldsymbol{\beta}_m, \boldsymbol{\Sigma}_i\right] = \sum_{s=1}^{1,000} \mathbb{1}\left(\tilde{\theta}_{mkis} \geq \underline{\Theta}_m(\tilde{\theta}_{pkis}; \pi_0^i, \pi_1^i, \widehat{\gamma}_c)\right)/1{,}000$, but this basic approach would hamper numerical optimization by introducing 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 (w/density levels/slopes equalling 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. Details on our numerical approach are in Online Appendix B.4. Computation time averaged 12-36 hours per Tobit specification on a Windows computer with a 12^{th} generation Intel processor (18 physical cores, 24 logical) and 64GB RAM, running MATLAB version 2022b.

These variables are summarized in Table 8 (Appendix A.3). For some of our survey covariates we have small fractions of observations missing, between 5% and 9% depending on which variable (see Table 8). 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.[32] Finally, as a robustness check on our primary Tobit specification, we add a final set of covariates derived from a parent survey. These include *Involved Parent*, a dummy for whether self-reported average time spent with the child on daily schoolwork weakly exceeds 2 hours; *BigFamily*, a dummy for 3 children or more living in the household; and two birth-order covariates, *Middle Child* (i.e., at least one younger and one older sibling), and *Youngest Child* (i.e., at least one older and no younger siblings); the omitted category is oldest child status (including only children). The main shortcoming of the parent survey data are that they are only available for roughly 20% of the sample (see Table 9, Appendix A.3). When incorporating these variables, we use the same imputation techniques to replace missing values for the rest of the sample. Assuming that data are essentially missing at random, conditional on all other available covariates, this approach limits the statistical power of the parent survey variables—the larger the sub-sample where they are available, the more power—rather than introducing bias.

5.1. **Empirical Results: Student Type Decomposition.** Tables 2 and 3 report results from Tobit regressions of student types on covariates. Recall that both equations for $\log(\theta_p)$ and $\log(\theta_m)$ are jointly estimated, which is why the log-likelihood function values for similarly numbered specifications across Tables 2 and 3 are the same. Specification (1) includes only gender, race, age, and neighborhood-SES controls, while (2) introduces school effects, and the remaining columns add increasing sets of additional variables. Specification (4) includes all controls that are available for the full sample, and (5) provides a final robustness check by adding the parent survey controls which were available for a 20% sub-sample.

5.1.1. *Mean Shifter Results.* Table 3 speaks to external validity of our motivation index θ_m derived from revealed preference under our experimental incentives. Recall that a potential limitation of our approach is that it may measure only willingness to allocate *extra time* to math learning, beyond regular schoolwork. Do raw $\hat{\theta}_m$ estimates represent deeper motivational differences across students, or do they reflect differences in baseline coursework load and differing levels of academic burnout? Our student survey data allow us to directly test this hypothesis: if it is true then the coefficient on *Reg Study Time* in Table 3 should be positive. This would be consistent with the idea that students with low measured motivation

[32] To fix ideas, suppose that the set of covariates includes $\boldsymbol{X} = \{1, X_{1i}, X_{2i}, X_{3i}\}_{i=1}^{N}$, but the value of X_{1i} is missing for some i. Imputation proceeds in two steps. First, we 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 $\hat{X}_{1i} = [1, X_{2i}, X_{3i}]\hat{\boldsymbol{\rho}}$ for each i with (only) X_{1i} missing.

TABLE 2. TOBIT REGRESSION RESULTS: PRODUCTIVITY

DEP VAR:	**(1)**		**(2)**		**(3)**		**(4)**		**(5)**	
$\boldsymbol{log\,(\theta_p)}$	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 Fixed Effects (District 1 Omitted)										
Joint Exclusion P-Value ($df=2$)**:**			$<10^{-16}$		$<10^{-16}$		2.2×10^{-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		0.444	

Notes: ***Higher* dependent variable values** $\log(\theta_p)$ **imply *lower* study-time productivity. 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-R^2** is $1 - \mathrm{E}[Var(\hat{\eta}_{pi})]/\left(Var(\boldsymbol{X}_{pi}\hat{\beta}_p) + \mathrm{E}[Var(\hat{\eta}_{pi})]\right)$.

(i.e., high $\log(\theta_m)$) are merely those who are more committed for regular coursework duties.[33] However, the coefficient on regular study time is actually *negative* and statistically indistinguishable from zero. This result lends credibility to our experimental/structural approach as tapping into latent factors that drive academic choices in students' everyday lives.

Although our model is primarily one of short-term choices, our pooling of 5$^{\text{th}}$and 6$^{\text{th}}$graders within the field experiment allows us to measure year-on-year evolution of types within the sample population. Coefficients on the *Grade-5* dummy in Tables 2 and 3 indicate that 5$^{\text{th}}$-graders and 6$^{\text{th}}$-graders are on average indistinguishable in their motivation for engaging in math activity, but 5$^{\text{th}}$-graders are less productive by 30% of a standard deviation, after controlling for the full set of student covariates. This estimate is quite stable across specifications in the productivity equation.

5.1.2. *School District Effects.* We now consider the role of school quality in shaping adolescent productivity and motivation. In particular, we wish to address two questions: after controlling for other contextual factors, (i) do school quality differentials make some kids

[33]Methodologically, even if this were the case, the researcher could remove the spurious apparent motivation from structural estimates by computing residual motivation, net of observed study time.

TABLE 3. TOBIT REGRESSION RESULTS: MOTIVATION

DEP VAR:	**(1)**		**(2)**		**(3)**		**(4)**		**(5)**	
$\boldsymbol{log\,(\theta_m)}$	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 Fixed 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	
Prnt Survey Cntrls (4)†	no		no		no		no		YES	
$\boldsymbol{N}$	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: ***Higher $log(\theta_m)$ values imply lower* willingness** to substitute time toward math activity and away from the best outside option. **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 $1 - \mathrm{E}[Var(\hat{\eta}_{mi})]/\left(Var(\boldsymbol{X}_{mi}\hat{\beta}_m) + \mathrm{E}[Var(\hat{\eta}_{mi})]\right)$.

more productive learners, and (*ii*) do they make some kids more motivated learners? In specification (4) of Table 2, after controlling for all student covariates, one's school enrollment predicts significant shifts in time required to complete learning tasks. From descriptive evidence in Table OS.3, one might guess that District 1's inputs—higher funding per student, larger fraction of budget devoted to instruction, more qualified and better paid faculty/administrators—are more advantageous to the student than District 2's and District

3's. Indeed, this expectation plays out in school value-added estimates in Table 2: switching from District 1 to District 2 or District 3 induces a reduction in a child's study-time productivity by 0.12 SD and 0.65 SD, respectively. School effects are highly jointly significant in both specifications (4) and (5), with individual school effects being statistically significant at the 5% level or less. They are also economically significant: the District-3 effect (relative to District 1) is roughly twice the gap between grade-5 and grade 6-students. Coefficient estimate magnitudes are remarkably stable across all four specifications where they appear, suggesting robustness of this result to inclusion of a rich set of other childhood contextual factors. On the other hand, estimated magnitudes of school effects on $\log(\theta_m)$ are much smaller, and not statistically different from zero, suggesting that school quality differentials play little role in driving motivation heterogeneity.

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 more academically adept students onto their rolls. 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 ($\theta_p, \theta_m$) (see Figure 11, Appendix A.3), they also appear to exert their own influence on productivity differentials as well. In Section 6 we investigate further channels through which school quality may operate, by shaping skill production technology, or the mapping between home study and durable skill gains.

5.1.3. *Racial differences.* In Tables 2 and 3 we see substantial correlations between both productivity and race within specification (1). Many well-meaning researchers and practitioners believe that many minority students from less affluent backgrounds, who often have worse academic outcomes, primarily struggle with a lack of motivation or engagement in academics. Our data rebut this idea; Tobit point estimates actually suggest the opposite, that Black and Hispanic students in our sample may be *more* conditionally motivated, though these effects are noisy and not statistically different from zero (race terms have a joint p-value of 0.30 in specification (5)). The idea of Black/Hispanic students being academically as motivated or more is also supported by patterns in our raw data: they are more likely to report math as their favorite subject than Whites/Asians (43.5% vs 30.2%, p-value 3.9×10^{-8}), and also exhibit higher levels of intrinsic motivation, on average (p-value 0.005).

Rather, in our data the main difference which could explain achievement gaps is a large racial productivity differential. It arises primarily because: (*i*) Black/Hispanic student enrollment is concentrated in Districts 2 and 3 (whereas White/Asian enrollment is concentrated in Districts 1 and 2), and (*ii*) there are residual factors proxied for by race, which persist after controlling for gender, school district, family structure, peer/adult support, neighborhood-SES, and home learning resources.[34] The conditional race gap in productivity is substantial,

[34]Other primary factors driving the productivity race gap are home learning resources—*no home internet* and *mobilefrac* (Blacks/Hispanics are 3.9× more likely to be without home internet, and logged 76% more page views from smartphones)—as well as *middle child* status, a proxy for parental resource scarcity.

with $\log(\theta_p)$ being nearly a full SD higher for Black students and 2/3 SD higher for Hispanics. This productivity gap dominates the estimated motivation advantage for Blacks/Hispanics, which explains why they completed fewer website tasks (Figure 9, Appendix A.3).

5.1.4. *Gender differences.* Figure 8 (Appendix A.3) suggests a moderate female advantage in motivation and a moderate female disadvantage in study-time productivity. In the final two specifications of Table 2, we find that learning productivity (i.e., rate of progress through learning tasks) is between 21% (spec. (4)) and 30% (spec. (5)) SD lower for females, relative to males. Although smaller than the racial gap, this is a non-negligible difference: the gender gap in learning productivity is between 2/3 and 100% of the gap associated with an extra year of schooling. In ongoing work, Cotton, et.al. (2024) present evidence to help explain the gender productivity gap: adolescent girls exhibit a systematically different approach to math problem solving, preferring more concrete, hand-written notes, relative to boys. On the other hand, in Table 3, we find that female adolescent students are also more motivated, with willingness to spend time on math study being 43% SD higher, on average, relative to males. In terms of total learning activity, the latter effect dominates, explaining why females complete more website learning tasks (see Figure 9 in Appendix A.3).

5.1.5. *Other considerations.* Psychic costs of effort have long been theorized within the education literature. We find suggestive evidence that (self-reported) preferences for math as a favorite subject raise motivation by 0.16 SD, while having math as one's least favorite reduces motivation by 0.16 SD. These estimates are somewhat noisy (joint p-value on subject preference parameters is 0.208), but the large magnitudes suggest a non-trivial role for psychic costs in children's self-investment choices. Perhaps our strongest result from Table 3 is that being *either* more extrinsically minded *or* more intrinsically minded are both strong indicators of responsiveness to external incentives for math study.[35] This forms part of a recent body of empirical work finding 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 result also contributes to the literature by directly quantifying another aspect of direct psychic costs: a one SD increase in intrinsic mindset score raises motivation by more than 1/3 SD, and is more than enough to overcome aversion to math as one's least favorite subject.

5.1.6. *Variances/Covariances of Productivity & Motivation Traits.* One advantage of the Tobit estimator is that it allows us to gauge the correlation between productivity and motivation for the population at large, including both active students (i.e., $A_i \geq 2$ on our website),

[35]For intrinsic/Extrinsic mindset scores, two questions each on the pre- and post-survey asked about a child'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 motivation scores. The presence of the fifth option means a student may be coded as exhibiting extrinsic mindset, intrinsic mindset, both, or neither.

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	

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). The table also presents 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}), of student types. The striking result from Table 4 is that in both specifications (4) and (5) the average type correlations are small and statistically indistinguishable from zero. Taking sampling variability into account, we can rule out moderate or large positive correlations between adolescent learners' productivity/motivation indices, and in fact, point estimates suggest a small to moderate *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 actually slightly 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" or "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 with the task of providing education services to a heterogeneous population: a child need not have it all to attain to a high degree of academic success. On the other hand, it also highlights salient challenges as well, suggesting that one-size-fits-all solutions may be ineffective when developing students who struggle primarily with low motivation vs. low productivity.

5.1.7. *Threats to Causal Interpretation of School Effect Estimates.* Our identification strategy for school value-added (VA) based on field-experimental data differs significantly 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, 2023)). Some studies use a similar many-schools approach in combination with some source of plausibly exogenous variation to tease apart selection on unobserved student traits from school VA; e.g., Dale and Krueger (2002) and Mountjoy and Hickman (2020). Other recent work by Abdulkadiroglu et al. (2020) combines standard VA methods with equilibrium theory to tease apart causal VA from preferences in a school-choice setting involving over 400 high schools in a major US city. 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 unobserved student traits, independently of school assignment. This, in turn, facilitates analysis of several forms of school VA in Sections 5 and 6.

To understand our VA identification strategy, it will be useful to contrast it with standard VA methods, such as those surveyed by Koedel et al. (2015). A typical approach would postulate some variant of a linear VA model like

$$Y_{ist} = \beta_0 + Y_{ist-1}\beta_1 + \boldsymbol{X}_{ist}\boldsymbol{\beta}_2 + \epsilon_{ist}, \quad \epsilon_{ist} = \xi_i + \zeta_s + e_{ist}, \tag{9}$$

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 equations (6) and (9) differ: ξ_i is the effect of unobserved student ability on academic outcome Y_{ist}, while residuals (η_p, η_m) are unexplained components of unobserved student ability itself.

Thus, the primary considerations in attaching a causal interpretation to school effect estimates in Tables 2 and 3 are two-fold: (i) controlling for factors that drive residential sorting patterns and also affect productivity/motivation, and (ii) controlling for other relevant sources of omitted variable bias. For (i), 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. In related research, Rothstein (2006) and Abdulkadiroglu et al. (2020) both 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 included in specifications (1)–(5) are adequately proxying for parents' residential sorting choices, and could explain why school effect estimates in Tables 2 and 3 are so robust to inclusion of additional covariates in specifications (3)–(5).

As for concern (ii)—general sources of omitted variable bias—as in the rest of the literature there is no way to completely rule this possibility out. Like other studies, the best we can do is control for what we believe are the most important sources of omitted variable bias as thoroughly as possible. This involved designing our data collection measures to gather as much relevant information as possible, given time and attention constraints involved in surveying adolescents during the math period of their school day. In order for the estimated school-district VA effects in Table 2 to be largely or mostly attributable to omitted variable

TABLE 5. DESCRIPTIVE STATISTICS: MATH EXAM SCORES BY SUB-SAMPLE

SUB-SAMPLE:	**ALL**	**FEMALE**	**MALE**	**BLACK**	**HISPANIC**	**WHITE/ ASIAN**
SIZE/FRACTION:	1,676	0.5078	0.4922	0.2691	0.1915	0.5394
Pre-Test Score, S:	13.40	12.71	14.11	7.93	7.94	18.07
(sample std. dev.)	*(8.96)*	*(8.23)*	*(9.62)*	*(6.13)*	*(6.10)*	*(8.35)*
ΔScore (All):	1.55	1.94	1.14	0.88	0.49	2.20
(sample std. dev.)	*(5.00)*	*(5.03)*	*(4.94)*	*(5.01)*	*(4.89)*	*(4.94)*
(p-value, H_0:No Change)	7.0×10^{-37}	2.4×10^{-29}	3.5×10^{-11}	1.9×10^{-4}	0.073	7.6×10^{-41}
ΔScore (Active Only):	2.67					
(sample std. dev.)	*(4.87)*					
(p-value, H_0:No Change)	8.0×10^{-51}					
ΔScore (Marg./Inactive Only):	0.51					
(sample std. dev.)	*(4.90)*					
(p-value, H_0:No Change)	0.0015					

Notes: italicized numbers in parentheses represent sample standard deviations. **The null hypothesis that website activity did not result in learning gains, or $H_0 : E[\Delta Score|Active] = E[\Delta Score|Marg./Inactive]$, is rejected by a two-sample t-test (p-value= $\mathbf{1.2\times10^{-17}}$).** 5th-graders are 47.3% of the sample, with 6th graders comprising the other 52.7%. Sub-sample proportions are close to that ratio for all gender and race groups.

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 a combination of race, gender, age, neighborhood socioeconomics, family/friend academic support, home learning resources, parental academic involvement, family size, and birth order.

6. EXPLORING THE DETERMINANTS OF MATH SKILL

This section uses structurally estimated student traits as a key input to recover the production technology of new math skill. Our approach highlights the inferential power to be had from directly quantifying latent motivation and productivity with field experimental data: they allow us to explicitly control for selection on unobserved student ability. We use student test scores from classroom pre- and post-tests to measure math skill (see Section 3.2.7). Let S_i denote i's initial math proficiency measured by pre-test score. Skill gains are thus changes in post-test scores at the end of the 3-week sample period, denoted ΔS_i. We allow student traits to not only determine short-term work choices, but also to influence the mapping between (T_i, A_i) and ΔS_i. This opens two additional channels through which school quality may operate: by altering "TFP"—i.e., by impacting skill formation that is (log-)separable from home-study—and by altering the rate at which child i converts a fixed volume of study activity into durable proficiency gains.

Table 5 shows descriptive statistics on average pre-test scores and proficiency gains by sub-group (see also Figures 7–8, Appendix A.3). 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; NAEP, 2019). On the pre-test, White/Asian students correctly answered nearly 10 additional questions (1.13 SD), on average, relative to Black/Hispanic students. The gender gap is relatively smaller, with the average male

correctly answering 1.4 more exam questions than the average female, or a 0.16 SD difference. The table also highlights how extracurricular math activity on our website during the sample period contributed to measured proficiency gains. Active students saw mean increases of 2.67 exam questions solved, while marginal/inactive students improved scores by only 0.51 points. Both changes are statistically significant at the 1% level.

6.1. **Determinants of mathematics proficiency.** Although our ultimate interest is in how home study activity (T_i, A_i) maps into math skill gains (ΔS_i), we begin with a preliminary investigation of initial proficiency (S_i), which we model as a Cobb-Douglass production function with student traits θ_{pi} and θ_{mi} as the primary inputs. Production shares and total factor productivity (TFP) terms are allowed to vary by individual i:[36]

$$S_i = TFP_i \times \theta_{pi}^{\alpha_{pi}} \times \theta_{mi}^{\alpha_{mi}} \times \epsilon_i, \quad TFP_i > 0, \;\; \alpha_{pi} < 0, \;\; \alpha_{mi} < 0. \tag{10}$$

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

$$\log(TFP_i) = \boldsymbol{W}_i\boldsymbol{\alpha}_0, \quad \alpha_{pi} = \boldsymbol{W}_i\boldsymbol{\alpha}_p, \quad \text{and} \quad \alpha_{mi} = \boldsymbol{W}_i\boldsymbol{\alpha}_m, \tag{11}$$

with $\boldsymbol{W}_i$, including various student-level contextual factors. The error term ϵ_i is an idiosyncratic shock that accounts for cumulative impacts of transitory shocks to HC production, and noise in the exam instrument used to measure math skill.

A student's pre-test score, S, provides a baseline measure of skill stock, while productivity, θ_p, governs the rate at which learning tasks are traversed during the process of augmenting skill stock. While the two concepts are related, they are not the same. Initial proficiency stock S is a measure of a child's ability to correctly solve math problems in a controlled, timed, classroom environment, without any real-time feedback or access to external aids. Productivity θ_p measures the time needed to correctly solve math problems in an un-structured homework setting, given real-time feedback on incorrect answers, and access to textbooks, examples, notes, and assistance from friends or family.

6.1.1. *Estimating the model.* Substituting (11) into (10), the initial proficiency model is equivalent to a regression of $\log(S_i)$ on $\log(\theta_{pi})$, $\log(\theta_{mi})$, and $\boldsymbol{W}_i$, with pair-wise interactions:

$$\log(S_i) = \boldsymbol{W}_i\boldsymbol{\alpha}_0 + \boldsymbol{W}_i\boldsymbol{\alpha}_p \log(\theta_{pi}) + \boldsymbol{W}_i\boldsymbol{\alpha}_m \log(\theta_{mi}) + \log(\epsilon_i). \tag{12}$$

In our full specification, the covariate vector $\boldsymbol{W}_i$ contains a constant and 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; home resource proxies (*no home internet*, *mobilefrac*, *tabletfrac*); and parent survey controls (*Involved Parent*, *BigFam*, *MiddleChild*, and *YoungestChild*). Note that each of these factors may have a direct impact (through the TFP terms $\boldsymbol{W}_i\boldsymbol{\alpha}_0$), and an indirect impact (through the slope terms $\boldsymbol{W}_i\boldsymbol{\alpha}_p$ and $\boldsymbol{W}_i\boldsymbol{\alpha}_m$). In order to interpret school effects causally, we require the following:

[36]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 *good* thing for skill development.

Assumption 7. $E[\boldsymbol{W}_i^\top \log(\epsilon_i)|\theta_{pi}, \theta_{mi}] = \mathbf{0}$.

Assumption 7 highlights the advantages of our school VA approach based on structural type estimates. Among our analyses in Sections 5 and 6, the initial proficiency model (12) is most comparable to the standard VA framework (equation (9)), but with three important changes. First, 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 (contained in $\boldsymbol{W}_i$). Second, we need not assume latent ability is a single index; rather, we can model it as $\xi_i = (\theta_{pi}, \theta_{mi})$, with different student traits potentially having distinct effects on the outcome. Third, we can capture the direct impact of school quality through its contribution to the $\boldsymbol{W}_i\boldsymbol{\alpha}_0$ (TFP) term, while allowing for *interactions* between school quality and student quality through its contributions to the $\boldsymbol{W}_i\boldsymbol{\alpha}_p$ and $\boldsymbol{W}_i\boldsymbol{\alpha}_m$ (production share) terms.

There are three implementation challenges to overcome. First, we only have structural point estimates of $(\theta_{pi}, \theta_{mi})$ for active students ($A_i \geq 2$), so we have a missing variables problem in equation (12). This is straightforward to address using Tobit projections from the previous section: for marginal/inactive students ($A_i<2$) we use conditional expectations,[37]

$$\left(\widehat{\log(\theta)_{pi}}, \widehat{\log(\theta)_{mi}}\right) = E\Big[\big(\log(\theta_p), \log(\theta_m)\big)\Big|\boldsymbol{X}_{pi},\ \boldsymbol{X}_{mi},\ A_i<2,\ \pi_{1i};\ \widehat{\boldsymbol{\beta}}_p, \widehat{\boldsymbol{\beta}}_m, \widehat{\boldsymbol{\Sigma}}_i\Big].$$

Specifications (1)–(4) in Table 6 are based on projections from Tobit specification (4). Specification (5) in Table 6, which adds parent survey variables, is based on projections from Tobit specification (5), which also included parent survey information.

The second challenge is an errors-in-variables problem stemming from sampling variability in student fixed-effect estimates. We compute Empirical Bayes (EB) estimates of (θ_p, θ_m) to reduce attenuation bias by shrinking each fixed effect 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. Morrix, 1983; Abdulkadiroglu et al., 2020) is to assume a normal prior over the true fixed effect, $\log(\theta_{ji})$, and the estimation residual, r_{ji} for $j=p,m$. This implies a shrinkage factor of $\lambda_{ji}=\nu_j^2/\left(\nu_j^2+\nu_{rji}^2\right)$, where ν_j^2 is the estimated variance of true $\log(\theta_{ji})$, and ν_{rji}^2 is the estimated sampling residual variance on $\widehat{\log(\theta_{ji})}$ for individual i's trait $j=p,m$.[38] This results in the following EB estimates for student traits to be used as regressors in (12): $\log(\theta_{ji})_{EB}=\lambda_{ji}\widehat{\log(\theta_{ji})} + (1-\lambda_{ji})\frac{\sum_{i=1}^N \widehat{\log(\theta_{ji})}}{N}$, $j=p,m$.

Third, our unbalanced panel data implies that the error terms in equation (12) may exhibit heteroskedasticity. Formal tests reveal that the null hypothesis of homoskedastic errors is indeed rejected in all specifications. Therefore, we estimate the production parameters via feasible generalized least squares in the familiar way, as outlined in Wooldridge (2016), and we use heteroskedasticity-robust standard errors for inference.

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

[38] An alternative approach is to restrict the shrinkage forecast of $\log(\theta_{ji})$, given $\widehat{\log(\theta_{ji})}$, to linear projections (e.g. Chetty et al., 2014), which implies the same shrinkage factor λ_{ji}.

TABLE 6. INITIAL MATH PROFICIENCY (Cobb-Douglas)

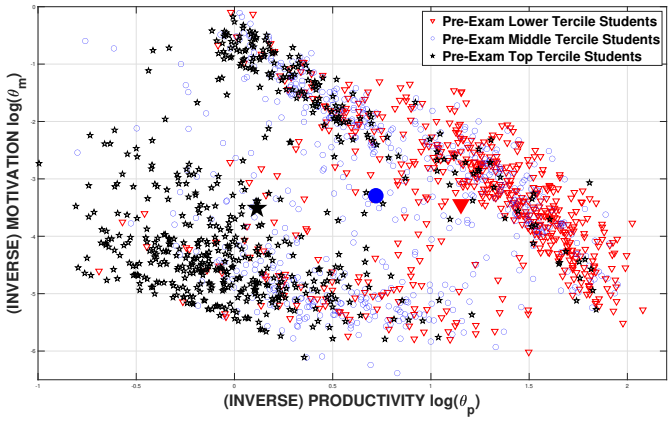
SPEC: **DEP VAR: log (S_1)**	**(1)** **(Mean; SD)**	**(2)** **(Mean; SD)**	**(3)** **(Mean; SD)**	**(4)** **(Mean; SD)**	**(5)** **(Mean; SD)**
$(\log(\widehat{TFP}_i))$	(3.207; 0)	(3.076; 0.183)	(3.033; 0.205)	(3.011; 0.233)	(3.023; 0.263)
θ_p **Production Share** $(\widehat{\alpha}_{pi})$	(−0.397; 0)	(−0.305; 0.120)	(−0.269; 0.112)	(−0.262; 0.116)	(−0.248; 0.133)
θ_m **Production Share** $(\widehat{\alpha}_{mi})$	(−0.024; 0)	(−0.027; 0.008)	(−0.033; 0.019)	(−0.037; 0.026)	(−0.032; 0.030)
	Avg SD Effect	**Avg SD Effect**	**Avg SD Effect**	**Avg SD Effect**	**Avg SD Effect**
log(*TFP*)	N/A	0.4348***	0.4869***	0.5551***	0.6269***
(joint p-value)		$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$
log(θ_p)	-0.6676***	-0.5124***	-0.4516***	-0.4399***	-0.4043***
(joint p-value)	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$	$(< 10^{-16})$
log(θ_m)	-0.0693***	-0.0788***	-0.0966***	-0.1077***	-0.1063***
(joint p-value)	(1.5×10^{-4})	(1.6×10^{-5})	(1.4×10^{-8})	(1.5×10^{-4})	(3.3×10^{-8})
INDIVIDUAL CONTROL VARIABLES:					
Grade 5 $(\widehat{\alpha}_{03}, \widehat{\alpha}_{p3}, \widehat{\alpha}_{m3})$	—	—	-0.2188***	-0.2028***	-0.2260***
(joint p-value)			(3.0×10^{-10})	(4.4×10^{-9})	(7.2×10^{-10})
Female $(\widehat{\alpha}_{04}, \widehat{\alpha}_{p4}, \widehat{\alpha}_{m4})$	—	—	-0.0549***	-0.0889***	-0.1846***
(joint p-value)			(1.5×10^{-4})	*(0.001)*	(6.2×10^{-5})
Black $(\widehat{\alpha}_{05}, \widehat{\alpha}_{p5}, \widehat{\alpha}_{m5})$	—	—	-0.1468***	-0.1222***	-0.1122***
(joint p-value)			*(0.007)*	*(0.007)*	*(0.009)*
Hispanic $(\widehat{\alpha}_{06}, \widehat{\alpha}_{p6}, \widehat{\alpha}_{m6})$	—	—	0.0229*	0.0114**	0.1136***
(joint p-value)			*(0.073)*	*(0.021)*	*(0.003)*
SCHOOL EFFECTS (District 1 omitted):					
Joint P-Value, All Terms ($df = 6$)**:**		$(< 10^{-16})$	(3.1×10^{-16})	(4.1×10^{-5})	(6.3×10^{-6})
Joint P-Value, TFP Only ($df = 2$)**:**		(8.8×10^{-10})	(9.7×10^{-7})	*(0.136)*	*(0.222)*
Joint P-Value, Slopes Only ($df = 4$)**:**		(6.2×10^{-15})	*(0.122)*	*(0.728)*	*(0.454)*
District 2 $(\widehat{\alpha}_{01}, \widehat{\alpha}_{p1}, \widehat{\alpha}_{m1})$	—	-0.3138***	-0.3111***	-0.2390***	-0.1105**
(joint p-value)		$(< 10^{-16})$	(1.3×10^{-8})	*(0.002)*	*(0.047)*
District 3 $(\widehat{\alpha}_{02}, \widehat{\alpha}_{p2}, \widehat{\alpha}_{m2})$	—	-0.6747***	-0.7343***	-0.6505***	-0.6897***
(joint p-value)		$(< 10^{-16})$	(3.2×10^{-11})	(4.4×10^{-5})	(1.4×10^{-5})
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	0.423	0.494	0.516	0.523	0.535

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 $log(S_1)$, of 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.

6.1.2. *Empirical Results.* Table 6 presents estimates of equation (12). For ease of interpretation given numerous variable interactions, the table reports *Average SD Effects*, defined as mean induced shift in standard deviations of $\log(S)$ from an increase in a continuous control variable of one standard deviation, or a 0-to-1 change for binary controls. SD effects encapsulate influence through all channels, both direct and indirect.

Table 6 results produce several interesting insights. First, both θ_p and θ_m are significant determinants of initial math skill, but productivity θ_p plays a clearly dominant role between

FIGURE 5. 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 (5)). Large bolded shapes are within-group means.

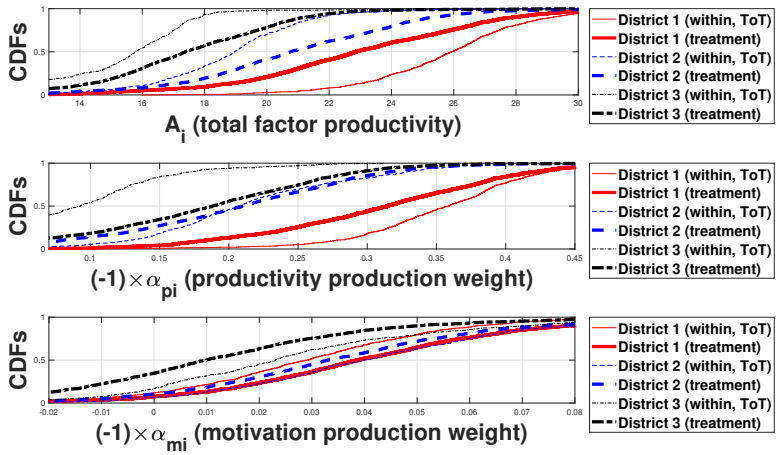
the two with a SD effect that is 4.1–3.8 times as big in specifications (4) and (5). Figure 5, a scatter-plot of $(\log(\theta_{pi}), \log(\theta_{mi}))$, further illustrates this point. The figure separates students by pre-test score terciles—triangles for the lower tercile, circles for the middle, and stars for the upper—and we include large shaded shapes to mark the average motivation and productivity values within each performance tercile.

Structural fixed-effect estimates depict wide variation in productivity-motivation pairs across students, but a comparison of the means reveals little difference in average motivation across the three terciles. Both results support the same conclusion: *lower performing students are not predominantly less motivated than higher performing students*. This is true regardless of whether performance is judged on the completion of website learning tasks (Figure 1) or proficiency assessment scores (Table 6 and Figure 5). Rather, the strongest distinguishing trait is learning productivity: students in higher exam terciles require substantially less time to complete homework, allowing them to traverse more learning task volume before opportunity costs rise too high to rationally continue study.

These results 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 (such as those studied in Fryer (2011); Levitt et al. (2016)).[39] These groups are either more or no-less willing to devote time to study than their peers, suggesting that motivation is not the primary barrier limiting their progress. Moreover, Table 6 suggests that, since TFP is about 5.2–5.9 times as important as θ_m, and θ_p is 4.1–3.8

[39]Gneezy et. al. (2019) also adds important insights for inducing effort on one-off tests.

FIGURE 6. Cobb-Douglas Parameters by School District



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*.

times as important, efforts to further incentivize effort by students from marginal groups may struggle to overcome the more salient productivity disadvantages they face.[40]

The second insight from Table 6 is evidence that school quality influences skill production technology in important ways. The magnitudes of the school district effects again strongly conform to the pattern one might suspect from the suggestive evidence in Table OS.3: Switching from District 1 (the high performing district) to District 2 (the middling school district) or District 3 (the struggling school district) entails substantially less effective skill production technology. We also see suggestive evidence that VA differences across school districts may not be one of *levels* only, but of the *shapes* of the production technologies.[41] Figure 6 plots empirical CDFs of student-specific production parameters and illustrates this idea: higher VA schools appear to have better TFP and also to get more production from student characteristics as well.

Third, we also 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* students in the sample). This means that the extra benefit in math skill development from improving a student's underlying characteristics declines as those characteristics become more favorable.

[40]This may explain why conditional cash transfers to students or families for improved academic performance have often resulted in limited returns (e.g., Fryer, 2011). Levitt et al. (2016) find limited returns to such conditional transfers in Chicago-area schools, which is the setting of our experiment. Leuven et al. (2010) show evidence among university students that those who are already performing well tend to respond most to financial incentives. C. Cotton, Nordstrom, Nanowski, and Richert (2024) find that significant effects on academic progress from information interventions may come at prohibitive costs. In ongoing work, C. S. Cotton, Hickman, List, and Sun (2024) explore structural market-design of academic incentives.

[41]Table 6 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).

This also implies that the marginal value of investments which may influence study productivity (e.g., tutors, improved educational resources, etc.) is higher for children with less advantageous productivity traits θ_{pi}, which is in line with other recent results by Agostinelli and Wiswall (2023), among others.

6.2. **Analysis of Study Effort and Proficiency Gains.** The previous section estimated a reduced-form production technology for initial proficiency stock. With our field experimental design we can go a step further by incorporating data on interim math learning activity and exam score shifts over the sample period. We model incremental proficiency gains as a quadratic complete polynomial in study time T_i and learning task volume A_i:

$$\Delta S_i = \Delta_{0i} + \Delta_{1i} T_i + \Delta_{2i} T_i^2 + \Delta_{3i} A_i + \Delta_{4i} A_i^2 + \Delta_{5i} (T_i \times A_i) + \varepsilon_i. \tag{13}$$

Once again, regression parameters depend on student covariates, with $\Delta_{ji} \equiv \boldsymbol{V}_i \boldsymbol{\delta}_j$ for $j = 0, 1, ..., 5$, being a single index of covariate vector $\boldsymbol{V}_i = [\boldsymbol{W}_i, S_i, \log(\theta_{pi}), \log(\theta_{mi})]$ for student i, including $\boldsymbol{W}_i$, with initial proficiency and student traits as additional controls. By including student types $\log(\theta_{pi})$ and $\log(\theta_{mi})$ in $\boldsymbol{V}_i$, we allow them to play a dual role in shaping skill acquisition: aside from driving choices (T_i, A_i) they may alter the rate at which a fixed volume of study activity is converted into durable skill gains. Including initial proficiency S_i as a control allows for possible decreasing-returns-to-scale technology, where proficiency score gains of a fixed size become more difficult as a student achieves greater subject mastery. In order to causally interpret school effects we require the following condition:

Assumption 8. $E[\boldsymbol{V}_i^\top \varepsilon_i | \theta_{pi}, \theta_{mi}] = \mathbf{0}$.

Assumption 8 once again highlights the advantages of our school VA approach based on structural type estimates. The incremental gains model (13) takes an additional step forward from standard VA models (equation (9)), by incorporating information on tracked home-study activity, and mapping these onto short-run measured changes in the outcome variable. This allows for a richer formulation of the production technology, and provides a novel window into the micro-foundations of learning. It also retains the advantages of the reduced-form VA model of the previous section: we are able to bring latent student ability out of the error term and explicitly control for selection on unobservables; we can still model ability ξ_i as a two-dimensional object $(\theta_{pi}, \theta_{mi})$; and we can still capture the direct impact of school quality through its contribution to the Δ_{0i} ("TFP") term, while allowing for *interactions* between school quality and student home study through its contributions to the Δ_{ji} (slope) terms, $j = 1, \ldots, 5$. While the school interactions included in (13) are somewhat different from those in (12)—school dummies interacted with *choices*, $\{T, T^2, A, A^2, T{\times}A\}$, versus school dummies interacted with *types*, $\{\theta_p, \theta_m\}$—it is important to remember that (T_i, A_i) are functions of $(\theta_{pi}, \theta_{mi})$, as formalized by the student choice model in Section 2. Thus, the interaction terms retain the same spirit, and moreover, the two models of skill technology are mutually consistent.

6.2.1. *Estimation.* The empirical strategy here faces similar challenges as in Section 6.1, and we employ similar coping strategies: EB shrunk type forecasts and FGLS estimation with heteroskedasticity-robust standard errors. Results are summarized in Table 7, again using *Average SD Effects* rather than reporting long lists of (up to 116) parameter estimates. For reference, recall that one SD of skill gains ΔS is 5 exam-score points, or 5 extra problems solved correctly out of 36, with a 40-minute time limit.

6.2.2. *Empirical Results.* While interpreting SD Effect results from Table 7, the reader should remember that they involve many complicated interactions between various factors, and are therefore quite heterogeneous across different students with a diverse set of life circumstances, including different schools, skill stocks, latent traits, homes, ages, genders, and racial backgrounds. Baseline learning $\widehat{\Delta}_{0i}$—that occurred independently of extracurricular website math activity—has a mean of 0.71 (0.77) exam-score points in specification (4) (specification (5)) over the 3-week sample period. Baseline learning also varied broadly across different students and contexts, with a standard deviation of 1.4 (1.6) exam-score points.

When interpreting SD Effects for the primary productive inputs T_i and A_i, it is not a well-posed thought experiment to hold one fixed while varying the other, since learning tasks require student time. Rather, simultaneous SD increases in T_i and A_i imply an average net increase of between 1.17(=2.01-0.84, specification (4)) and 0.57(=1.07-0.50, specification (5)) SD of skill gain ΔS. This translates into an improvement range of 2.85–5.85 exam-score points, or between 25 and 51 practice problems solved (26–53 minutes practice time) per exam point gained, for the average child. Learning task completion A_i is the primary driver of skill development; idiosyncratic SD Effects of T_i are *negative* for roughly 4/5 of the student sample, meaning that total time spent actually tempers (but doesn't negate) the conversion rate of task completion A_i into new durable skill. This result points to some important policy-relevant insights: adolescent math proficiency is developed through experience in correctly solving a wide variety of problems, rather than by spending more time thoroughly understanding solutions to fewer problems. It also further highlights the importance of addressing productivity differentials for struggling students.

Table 7 results suggest that, holding productivity type θ_p fixed, as children progress from 5^{th}-grade to 6^{th}-grade they become more effective at learning, with the total year-on-year difference being over 1/5 SD of ΔS, through changes to both slopes and the intercept. In other words, while students gain more experience as learners, they not only become more adept at subject matter, but they also become more adept at the act of learning itself. We also find strong evidence that one's productivity type θ_p alters the shape of the learning technology in an economically meaningful way, by influencing both slopes and intercepts. Thus, students who are more productive with their learning time also tend to derive more durable skills from a fixed volume of learning tasks as well. Our grade-5 and $\log(\theta_p)$ effects from Table 7, along with the grade-5 effect from Table 2, corroborate similar findings by Cunha et al. (2010) on the complementarity of skill development over time.

TABLE 7. PRODUCTION OF INCREMENTAL GAINS IN MATH SKILL

SPEC:	**(1)** **(Mean; SD)**	**(2)** **(Mean; SD)**	**(3)** **(Mean; SD)**	**(4)** **(Mean; SD)**	**(5)** **(Mean; SD)**
DEP VAR: ΔS	(1.549; 5.003)	(1.549; 5.003)	(1.549; 5.003)	(1.549; 5.003)	(1.549; 5.003)
Baseline Learning $\widehat{\Delta}_{0i}$	(0.511; 0)	(0.831; 1.272)	(0.714; 1.355)	(0.706; 1.378)	(0.770; 1.577)
	Avg SD Effect	**Avg SD Effect**	**Avg SD Effect**	**Avg SD Effect**	**Avg SD Effect**
Baseline Learning $\widehat{\Delta}_{0i}$ *(joint p-value)*	—	0.2543*** (1.4×10^{-14})	0.2709*** (4.6×10^{-14})	0.2754*** (8.0×10^{-11})	0.3153*** (1.6×10^{-14})
$\boldsymbol{T}$ $\left(\widehat{\Delta}_{1i},\widehat{\Delta}_{2i},\widehat{\Delta}_{5i}\right)$ *(joint p-value)*	0.1477*** (4.1×10^{-5})	-0.5024*** $(<10^{-16})$	-0.6222*** $(<10^{-16})$	-0.8383*** $(<10^{-16})$	-0.5038*** $(<10^{-16})$
$\boldsymbol{A}$ $\left(\widehat{\Delta}_{3i},\widehat{\Delta}_{4i},\widehat{\Delta}_{5i}\right)$ *(joint p-value)*	0.3196*** *(0.001)*	1.2424*** $(<10^{-16})$	1.5160*** $(<10^{-16})$	2.0135*** $(<10^{-16})$	1.0660*** $(<10^{-16})$
$\boldsymbol{S}$ $(\widehat{\delta}_{0,1},\ldots,\widehat{\delta}_{5,1})$ *(joint p-value)*	—	-0.4194*** $(<10^{-16})$	-0.4334*** $(<10^{-16})$	-0.4166*** $(<10^{-16})$	-0.4539*** $(<10^{-16})$
$\mathbf{log}(\boldsymbol{\theta_p})$ $(\widehat{\delta}_{0,2},\ldots,\widehat{\delta}_{5,2})$ *(joint p-value)*	—	-0.3423*** (7.5×10^{-4})	-0.2787** *(0.006)*	-0.4202*** (8.4×10^{-9})	-0.5016*** (1.8×10^{-10})
$\mathbf{log}(\boldsymbol{\theta_m})$ $(\widehat{\delta}_{0,3},\ldots,\widehat{\delta}_{5,3})$ *(joint p-value)*	—	-0.0488* *(0.099)*	0.0271 *(0.110)*	0.0457 *(0.195)*	0.0690* *(0.065)*
Grade 5 $(\widehat{\delta}_{0,4},\ldots,\widehat{\delta}_{5,4})$ *(joint p-value)*	—	—	-0.2221*** (2.4×10^{-4})	-0.2329*** (9.2×10^{-7})	-0.2192*** (9.2×10^{-7})
Female $(\widehat{\delta}_{0,5},\ldots,\widehat{\delta}_{5,5})$ *(joint p-value)*	—	—	0.0460 *(0.336)*	0.1335* *(0.068)*	0.1010 *(0.1685)*
Black $(\widehat{\delta}_{0,6},\ldots,\widehat{\delta}_{5,6})$ *(joint p-value)*	—	—	0.1287*** (3.5×10^{-7})	0.1476*** (1.3×10^{-5})	0.2118*** (3.3×10^{-15})
Hispanic $(\widehat{\delta}_{0,7},\ldots,\widehat{\delta}_{5,7})$ *(joint p-value)*	—	—	0.0636*** *(0.002)*	0.0841*** *(0.003)*	0.1285** *(0.009)*
SCHOOL EFFECTS (District 1 omitted):					
Joint P-Value, All Terms $(df=12)$**:**		$(<10^{-16})$	$(<10^{-16})$	(3.0×10^{-4})	(2.4×10^{-11})
Joint P-Value, Intercepts Only $(df=2)$**:**		*(0.055)*	*(0.006)*	*(0.462)*	*(0.179)*
Joint P-Value, Slopes Only $(df=10)$**:**		$(<10^{-16})$	$(<10^{-16})$	*(0.001)*	(1.0×10^{-6})
District 2 $(\widehat{\delta}_{0,8},\ldots,\widehat{\delta}_{5,8})$ *(joint p-value)*	—	-0.1516*** (2.6×10^{-4})	-0.2229*** (4.9×10^{-6})	-0.1084*** *(0.009)*	-0.0947*** (2.2×10^{-12})
District 3 $(\widehat{\delta}_{0,9},\ldots,\widehat{\delta}_{5,9})$ *(joint p-value)*	—	-0.4322*** $(<10^{-16})$	-0.5309*** $(<10^{-16})$	-0.4418*** (7.6×10^{-4})	-0.5518*** *(0.001)*
Nbhd-SES Cntrls (12)	no	no	no	YES***	YES***
Home Acad Support (6)	no	no	no	YES*	YES
Home Connectivity (18)	no	no	no	YES***	YES***
Prnt Survey Cntrls (20)	no	no	no	no	YES***
#Parameters	6	36	60	96	116
$\boldsymbol{N}$	1,494	1,494	1,494	1,494	1,494
$\boldsymbol{R^2}$	0.096	0.198	0.221	0.237	0.247

Notes: For context, $StDev(\Delta S)=5.00$ exam score points. **Avg SD Effect** measures total impact of a variable through the intercept Δ_{0i} (direct effect) *and* slope terms $\{\Delta_{1i},\ldots,\Delta_{5i}\}$ (interactions). It is the mean impact in standard deviations of ΔS (across all students) from switching a binary variable value from 0 to 1, or from increasing a continuous variable value by one standard deviation. Reported *joint p-values* are for the joint exclusion of all terms involving a given control. Significance at the 99%, 95% and 90% levels are denoted by ***, **, and *, resp. In spec. (4), interaction terms alone $(\widehat{\delta}_{1k},\ldots,\widehat{\delta}_{5k})$, $k=1,\ldots,7$ have the following *joint p-values*: 0.0127 for $\boldsymbol{S}$; 4.0×10^{-6} for $\mathbf{log}(\boldsymbol{\theta_p})$; 0.1490 for $\mathbf{log}(\boldsymbol{\theta_m})$; 2.0×10^{-5} for **Grade 5**; 0.1473 for **Female**; 0.0002 for **Black**; and 0.0122 for **Hispanic**. For parent survey controls we did not include $T\cdot A$ interactions to avoid numerical instability problems.

The table also gives further evidence of a decreasing returns to scale skill production technology: the Avg SD Effect of pre-test score S_i is significant (both statistically and economically) and *negative*. This implies that (over the short/medium run) as students reach

a higher level of mastery of $5^{\text{th}}/6^{\text{th}}$-grade math concepts, achieving further improvements of a fixed size in ΔS units requires increasingly more work.

Finally, we find once again that after controlling for the rich set of student covariates, school quality plays an important role in converting math activity into new skill stock. Moreover, the ordering among the three school districts is consistent with results from previous sections: all else equal, a switch from District 1 to District 2 or District 3 has a total effect, on average, of reducing skill augmentation by 0.11 SD and 0.44 SD (0.09 SD and 0.55 SD), respectively, under specification (4) (specification (5)). School district terms are highly significant all together, and individual District-1/District-2 total effects are still significant at the 1% level when taken separately. In contrast to the model of initial math proficiency, the data favor slopes as the main mechanism through which school quality impacts skill growth. In other words, holding student observed and unobserved traits fixed, more effective schools appear to act as a force multiplier for a child's home learning efforts.

6.3. **Threats to Causal Interpretation of School VA Effects on Skill Production.** Once again, our confidence in attaching causal interpretations to school VA estimates is bolstered by the stability of estimated magnitudes across model specifications. This is suggestive of having adequately controlled for relevant sources of omitted variable bias in Models (12) and (13). The core of our identification strategy in Section 6 has been the ability to directly control for selection on unobservables. Within the standard VA framework (equation (9)), this is equivalent to bringing latent student ability ξ_i out of the error term as an explicit regressor. To our knowledge, ours is the first paper to do so in estimating K-12 school VA.[42] Moreover, our structurally-motivated field experiment also allows us to model multiple dimensions of latent ability as $\log(\theta_{pi})$ and $\log(\theta_{mi})$. Of course, all claims to empirical causality rest to some extent on unverifiable assumptions. We require that remaining unexplained factors determining initial proficiency (Assumption 7) or incremental skill gains (Assumption 8) be conditionally uncorrelated with our full set of controls. In order for the school-district VA effects in Tables 6 and 7 to be largely or mostly attributable to omitted variable bias, there would have to exist some additional aspect of latent ability or other math-skill-relevant factor that is both strongly correlated with school assignment, and also *not* well proxied for by a combination of race, gender, age, neighborhood socioeconomics, family/friend academic support, home learning resources, parental academic involvement, family size, birth order, idiosyncratic study productivity θ_p, and idiosyncratic study motivation θ_m.

7. Conclusion

Since the 1960s, few areas in economics have grown as significantly or provided as many valuable insights as the study of human capital's role in economic growth and the examination of how education, learning, and skills are developed. A perusal of the popular press suggests

[42]Bodoh-Creed and Hickman (2024) use a related identification strategy to estimate college value-added in a Mincer equation for post-graduate income. They used observational college data and a model of Bayes-Nash competition for college admissions to derive a control function which proxied for 1-dimensional unobserved college-student types.

that most have accepted James Mill's dictum that "if education cannot do everything, there is hardly anything it cannot do." Yet, modern economies still strive to increase the number of their citizens completing higher education.

Gone are the days when societies can invest in only a small number of highly educated persons, where the primary goal of education is to pinpoint the few students who can succeed. Such systems historically invest a great deal more in the selection, rather than development, of students. These days, however, investment in the development of a broader set of students is important both for creating opportunities for the economic success and stability of individuals, and for innovation and growth within society. Quality education is no longer a luxury for a select few elite, but rather increasingly a necessity for anyone hoping to secure comfortable employment, let alone upward mobility.

This study adds to the vast literature by using a structurally-motivated field experiment to produce several important lessons for education policy. At the most fundamental level, we show that programs or policies that aim to close performance gaps by better motivating under-performing groups, either through information or incentives, may not be addressing the main barriers that constrain student performance. We show that under-performing students and groups, whether defined by race, gender, or school district, tend not to be any less motivated compared to their higher-performing peers. Rather, these under-performing students typically struggle to convert their study time and effort into learning task completion and proficiency gains. As such, effectively closing performance gaps likely requires something different than motivating under-performing students. Indeed, the effective closure of performance gaps should aim to improve their study-time productivity. This may mean improving access to high-quality education, tutoring, and supplemental learning resources, especially in early grades. It may also mean increasing the use of formative assessment and individualized curriculum, through teacher efforts or technology assisted learning.[43]

Our analysis also highlights key differences in quality across school districts, suggesting that students enrolled in less-affluent districts are at a substantial disadvantage compared to students in higher performing districts. Our approach makes it possible to bread this total effect down by several different mechanisms. First, after controlling for various contextual factors (residential sorting patterns, home background, etc.), more effective schools make their students more productive learners at home; that is they enable their students to traverse a critical mass of homework tasks before burning out. Second, more effective schools have a positive impact on learning independent of home-study activity, and third, they also influence the rate at which a child's study effort creates new durable math skill. Finally, since more

[43]These insights align with previous studies, which have found that education interventions focused on providing information or offering student incentives yield relatively small impacts on test scores (e.g., Baird, McIntosh, and Ozler (2021)). In contrast, programs that strengthen foundational math and literacy skills in early grades (e.g., Banerjee et al. (2016)), or tailor curriculum and teaching to meet individual learners' needs—through methods like tutoring, formative assessments, individualized education plans, or technology-assisted learning (e.g., Pitchford, Chigeda, and Hubber (2019), Outhwaite, Gulliford, and Pitchford (2017), Rodriguez-Segura (2020)) tend to have more substantial effects.

productive students also enjoy a higher conversion rate between effort and skill gains, school quality may indirectly strengthen this effect by bolstering productivity.

A key takeaway from the work of Heckman and colleagues, along with many others, is that investing in human capital yields a higher return than investing in physical capital. This insight suggests we should shift from an economy characterized by scarce educational opportunities to one that promotes and supports the development of all students over the life-cycle. A troubling observation about educational scarcity is that Black and Hispanic students report higher preferences for STEM but face significant disadvantages. They tend to be less affluent, often lack health insurance, and are mostly enrolled in schools with below-average budgets, faculty salaries, and teacher qualifications. Consequently, their standardized test scores lag behind those of White and Asian students, who generally have far better resource allocations. These facts together suggest adults are successfully advertising to Black and Hispanic children that STEM education is the way out of poverty. However, their communities, schools, and society at large are failing to follow up the marketing campaign by equipping them with the tools to effectively act on this perception.

Of course, any particular exercise leaves much on the sidelines. In our case, we should be clear that academic productivity and preferences may evolve over the long run. There is ample evidence (Bloom, 1964; Hunt, 1961) that academic efficiency can be modified by appropriate environmental conditions in the school and in the home. Factors such as the amount of time allowed for learning, quality of teacher or parent instruction, and the student's ability to understand instruction are important in determining the arc of learning alongside our studied characteristics. Indeed, they may serve as important complements. For example, an improvement in the quality of instruction yields important temporal returns: the student now must commit less time for learning the same amount of materials. Likewise, if the student lacks ability to understand teacher instruction (which could be due to low previous investment), the amount of time needed to learn increases. These are the dynamic complementarities that are a key aspect in the development of human capital (Cunha and Heckman (2007)). We reserve these discussions for another occasion but note that they are ripe for further theoretical and empirical inquiry.

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., & Wiswall, M. (2022). Estimating the technology of children's skill formation. *Journal of Political Economy*, *forthcoming*.

Agostinelli, F., & Wiswall, M. (2023). 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., Blackwell, D., & Girshick, A. (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica*, *17*, 213–244.

Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over time: Dynamic inconsistency in real effort tasks. *Quarterly Journal of Economics*, *130(3)*, 1067–1115.

Baird, S., McIntosh, C., & Ozler, B. (2021). Cash or condition? evidence from a cash transfer experiment. *Quarterly Journal of Economics*, *126*, 1709–1753.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., & Walton, M. (2016). Teaching at the right level: Evidence from randomized evalautions in India. *NBER Working Paper*. (wp 22746)

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.

Bettinger, E. (2012). Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics*, *94*, 686–698.

Bloom, B. S. (1964). *Stability and change in human characteristics.* New York: Wiley.

Bodoh-Creed, A., Hickman, B., List, J., Muir, I., & Sun, G. (2023). Stress testing a structural model of nonlinear pricing: Robust inference on intensive-margin consumer demand. *Working Paper, Washington University in St Louis Olin Business School*.

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*.

Buchholz, N., Shum, M., & Xu, H. (2023). Rethinking reference dependence: Wage dynamics and optimal taxi labor supply. *working paper, Princeton University Economics Dept.*.

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.

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(4)*, 855–902.

Chow, Y., & Robbins, H. (1963). On optimal stopping rules. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *2*, 33–49.

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. 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)

Cotton, C. S., Hickman, B. R., & List, J. A. (2024). Less proficient or more methodical: Insights on gender differences in math learning activity. *WashU-Olin working paper*.

Cotton, C. S., Hickman, B. R., List, J. A., & Sun, G. (2024). A market-design approach to reducing student achievement gaps: Insights for addressing heterogeneous productivity and motivation. *WashU-Olin working paper*.

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

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. (2019). 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.

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.

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.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132.

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. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica*, *68(3)*, 525–574.

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*.

Hamilton, B. H., Hickman, B. R., & Mohie, R. A. (2024). A new method for efficient computation of non-stationary dynamic programming problems with history dependence. *working paper, Washington University in St. Louis, Olin Business School*.

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.

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*.

Hotz, J., & Miller, R. (1993). Conditional choice probabilities and the estimation of dynamic models. *Review of economic Studies*, *60(3)*, 497–529.

Hunt, J. M. (1961). *Intelligence and experience*. New York: The Ronald Press Company.

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.

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.

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.

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. (2023). The determinants of teaching effectiveness: Evidence from a model of teachers' and students' interactions. *working paper,Olin Business School, Washington University in St Louis*.

Morrix, 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*.

NAEP. (2019). *National assessment of educational progress.* National Center for Education Statistics, Washington, D.C. (available online at http://nces.ed.gov/nationasreportcard/)

Outhwaite, L., Gulliford, A., & Pitchford, N. (2017). Closing the gap: efficacy of a tablet intervention to support the development of early mathematical skills in UK primary school children. *Computers & Education*, *108*, 43–58.

Pitchford, N., Chigeda, A., & Hubber, P. (2019). Interactive apps prevent gender discrepancies in early-grade mathematics in a low-income country in sub-Sahara Africa.

Developmental Science, *22*, e12864.

Rao, G. (2019). Familiarity does not breed contempt: Generosity, discrimination, and diversity in delhi schools. *American Economic Review*, *109(3)*, 774–809.

Rodriguez-Segura, D. (2020). Education technology in developing countries: A systematic review.

(EdPolicyWorks working paper)

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.

Snell, L. (1952). Applications of martingale system theorems. *Transactions of the American Mathematical Society*, *73(2)*, 293–312.

Torgovitsky, A. (2015). Identification of nonseparable models using instruments with small support. *Econometrica*, *83(3)*, 1185–1197.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16(2)*, 117–186.

Walls, J. (2017). *The glass castle*. Verago.

Wang, M.-T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review*, *33*, 304–340.

Westover, T. (2018). *Educated*. Random House.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*, 49–78.

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.

Appendix A. Observable Student Characteristics and Test Scores

A.1. **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. We obtained copies of 46 different standardized exams used by various U.S. states over the preceding decade, of which 30 were developed for 5^{th}graders and 16 were developed for 6^{th}graders.[44] The exams were then split into individual math problems, resulting in a bank of 370 unique grade-5 problems and 302 unique grade-6 problems.

[44]These state standardized math exams included the *California Standards Test* (2009), *Illinois Standards Achievement Test* (2003, 2006-2011, 2013), *Minnesota Comprehensive Assessments-Series III*, *New York State Testing Program* (2005-2010), *Ohio*

All 672 problems were pooled to expose both 5$^{\text{th}}$and 6$^{\text{th}}$graders to the same materials. This facilitated an even comparison between age groups, allowing us to cleanly estimate the effect of an additional year of schooling on skill formation.

We used Common Core Math Standards definitions to categorize each problem into one of 5 subject categories: (*i*) *equations and algebraic thinking*, (*ii*) *fractions, proportions, and ratios*, (*iii*) *geometry*, (*iv*) *measurement and probability*, and (*v*) *number system.*[45] For the pre-test and post-test, we randomly selected a large subset of problems from the math question bank and further categorized them as *easy*, *medium*, or *hard*, depending on their complexity level or number of steps required to solve. Finally, to ensure uniformity of subject content and difficulty level, both the pre-test and post-test were populated with similar sets of 36 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$^{\text{th}}$grade materials and the other 16 from 5$^{\text{th}}$grade materials, and the easy, medium, and hard categories were represented by 15, 12, and 9 questions respectively, spread evenly across each exam. We computed pre-test scores S_i and post-test scores S_{2i} by awarding one point for each correct answer, subtracting one quarter point for each incorrect answer (questions all had four possible choices), and neither adding nor subtracting points for answers left blank.

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 (and the scoring rule was explained in intuitive terms), 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 ≈160 (rounded to nearest 10 to preserve anonymity) US Census block groups where our test subjects resided, each of which can be thought of as a neighborhood. Within each neighborhood we collected mean household income (a proxy for affluence), and the fraction of minors with no private health insurance (a proxy for deprivation of non-school developmental resources).[46]

Achievement Test (2005), *State of Texas Assessments of Academic Readiness* (2011, 2013), *Texas Assessment of Knowledge and Skills* (2009), and *Wisconsin Knowledge and Concepts Examinations Criterion-Referenced Test* (2005).

[45]Common Core subject definitions for 5$^{\text{th}}$and 6$^{\text{th}}$grades (`http://www.corestandards.org/wp-content/uploads/Math` accessible as of September 2020) differ slightly; our 5-subject classification represents a merging of the two.

[46]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 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 8. DESCRIPTIVE STATISTICS: STUDENT COVARIATES BY SUB-SAMPLE

SUB-SAMPLE:	**ALL**	**FEMALE**	**MALE**	**BLACK**	**HISPANIC**	**WHITE/ASIAN**
SIZE/FRAC OF TOT:	1,676	0.5078	0.4922	0.2691	0.1915	0.5394
SCHOOL DISTRICT & NEIGHBORHOOD SOCIOECONOMICS						
Nbhd Mean Income ($1,000's)	$108.9	$108.9	$108.9	$80.8	$45.7	$132.0
(sample StDev, [fmo]*)*	*(41.5,*[0.05]*)*	*(41.1)*	*(41.9)*	(32.4)	*(23.2)*	*(24.6)*
Nbhd Uninsured Minors	0.252	0.253	0.252	0.378	0.616	0.072
(sample StDev, [fmo]*)*	*(0.297,*[0.06]*)*	*(0.297)*	*(0.297)*	*(0.293)*	*(0.231)*	*(0.129)*
District 1	0.465	0.475	0.455	0.007	0.044	0.843
(sample StDev, [fmo]*)*	*(0.499,*[0.00]*)*	*(0.499)*	*(0.498)*	*(0.083)*	*(0.205)*	*(0.364)*
District 2	0.268	0.260	0.276	0.650	0.103	0.136
(sample StDev, [fmo]*)*	*(0.443,*[0.00]*)*	*(0.439)*	*(0.447)*	*(0.477)*	*(0.304)*	*(0.343)*
District 3	0.267	0.266	0.269	0.344	0.854	0.021
(sample StDev, [fmo]*)*	*(0.442,*[0.00]*)*	*(0.442)*	*(0.443)*	*(0.475)*	*(0.353)*	*(0.143)*
FAMILY & RECREATIONAL TIME-USE VARIABLES						
#Adult Academic Helpers	1.140	1.163	1.117	1.128	0.615	1.328
(sample StDev, [fmo]*)*	*(0.848,*[0.07]*)*	*(0.821)*	*(0.875)*	*(0.892)*	*(0.724)*	*(0.789)*
#Peer Academic Helpers	0.789	0.907	0.666	0.852	0.887	0.728
(sample StDev, [fmo]*)*	*(0.783,*[0.07]*)*	*(0.792)*	*(0.756)*	*(0.825)*	*(0.766)*	*(0.765)*
#Gaming Systems at Home	1.570	1.474	1.660	1.648	1.480	1.554
(sample StDev, [fmo]*)*	*(1.135,*[0.00]*)*	*(1.130)*	*(1.133)*	*(1.299)*	*(1.096)*	*(1.056)*
Parental Permission for Video Gaming on Weekdays	0.878	0.882	0.874	0.809	0.888	0.909
(sample StDev, [fmo]*)*	*(0.327,*[0.00]*)*	*(0.322)*	*(0.332)*	*(0.393)*	*(0.316)*	*(0.287)*
Weekday Daily Recreational Internet Use (hrs)	1.766	1.790	1.740	1.908	1.788	1.694
(sample StDev, [fmo]*)*	*(1.201,*[0.09]*)*	*(1.166)*	*(1.236)*	*(1.290)*	*(1.210)*	*(1.150)*
Enrollment in Sports	0.669	0.639	0.700	0.548	0.455	0.807
(sample StDev, [fmo]*)*	*(0.471,*[0.00]*)*	*(0.481)*	*(0.458)*	*(0.498)*	*(0.499)*	*(0.395)*
Enrollment in Music	0.383	0.462	0.302	0.295	0.196	0.493
(sample StDev, [fmo]*)*	*(0.487,*[0.00]*)*	*(0.499)*	*(0.459)*	*(0.457)*	*(0.398)*	*(0.500)*
Enrollment in Clubs/ Other Activities	0.410	0.438	0.381	0.337	0.315	0.480
(sample StDev, [fmo]*)*	*(0.492,*[0.00]*)*	*(0.496)*	*(0.486)*	*(0.473)*	*(0.465)*	*(0.500)*
Fraction Leisure Time In Adult-Supervised Activity	0.351	0.356	0.345	0.317	0.274	0.392
(sample StDev, [fmo]*)*	*(0.172,*[0.05]*)*	*(0.172)*	*(0.171)*	*(0.167)*	*(0.181)*	*(0.158)*
ACADEMIC PREFERENCES & ATTITUDE VARIABLES						
Math Favorite Subj.	0.361	0.319	0.404	0.431	0.439	0.302
(sample StDev, [fmo]*)*	*(0.480,*[0.07]*)*	*(0.466)*	*(0.491)*	*(0.496)*	*(0.497)*	*(0.460)*
Math Least Favorite Subj.	0.216	0.254	0.176	0.277	0.212	0.189
(sample StDev, [fmo]*)*	*(0.411,*[0.08]*)*	*(0.435)*	*(0.381)*	*(0.448)*	*(0.410)*	*(0.392)*
Extrinsic Motiv. Score	0	-0.023	0.024	-0.222	-0.030	0.122
(sample StDev, [fmo]*)*	*(1,*[0.00]*)*	*(0.989)*	*(1.011)*	*(1.016)*	*(1.005)*	*(0.971)*
Intrinsic Motiv. Score	0	0.056	-0.058	0.010	0.150	-0.059
(sample StDev, [fmo]*)*	*(1,*[0.00]*)*	*(1.005)*	*(0.992)*	*(1.047)*	*(1.057)*	*(0.949)*

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 9. 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** and **Middle Child** are birth-order dummies for whether the child enrolled in the field experiment is the youngest (with at least one sibling) or middle child. The omitted birth-order category is **Oldest Child**, which includes only-child as a special case.

A.2. **Descriptive Statistics.** Table 8 presents descriptive statistics by demographic subgroup. In what follows, we adopt the terminology of referring to Blacks and Hispanics collectively as "under-represented minorities" or simply "minorities."[47]

On average, Black students in our sample live in neighborhoods with mean incomes moderately above that of the average student in Illinois ($71,602; see Online Appendix B), and Hispanic students in our sample live in neighborhoods with significantly lower mean incomes. White and Asian students in our sample live in neighborhoods with significantly higher incomes than the state average. The correlation between socioeconomics and race is also starkly apparent in uninsured minor rates, being higher among Blacks than Whites/Asians by a factor of 5.3, and higher among Hispanics by a factor of 8.6.

From survey responses we also see racial differences in terms of access to homework help, video game/internet usage, and participation in extracurricular activities. Whites/Asians have access to more adult academic helpers (including parents, grandparents, and tutors) and were more likely to be enrolled in sports and music. Black and Hispanic students are more likely to report that math is either their favorite or least favorite subject relative to their White/Asian peers. Minority students also self-reported higher levels of intrinsic motivation when completing school work, while White/Asian students are more likely to report being motivated by extrinsic factors such as satisfying parental or teacher expectations, or to earn a reward for satisfactory performance. Females in our sample also self-reported higher levels of intrinsic motivation, and lower levels of extrinsic motivation, relative to males.

[47]This convention follows the higher education literature, where Blacks and Hispanics are known to be proportionally under-represented at post-secondary education institutions. By contrast, Asian students, although a statistical demographic minority group, are proportionally over-represented at colleges generally, and particularly so at elite colleges, like their White counterparts. Thus, Asians do not satisfy the definition of a "URM" group. For ease of discussion, we will often refer to URMs as simply "minorities" for short, while recognizing this important caveat.

FIGURE 7. Distributions of Characteristics by Race

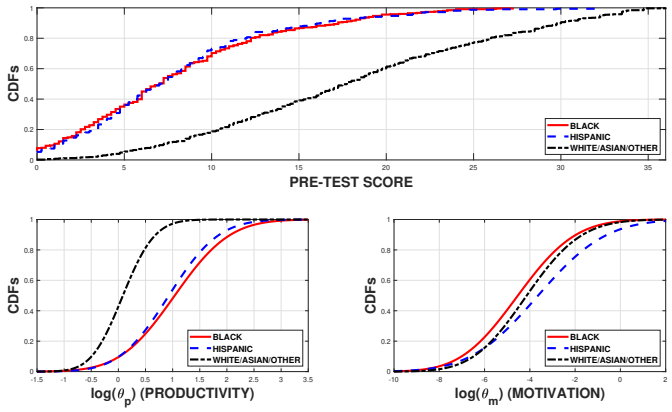


FIGURE 8. Distributions of Characteristics by Gender

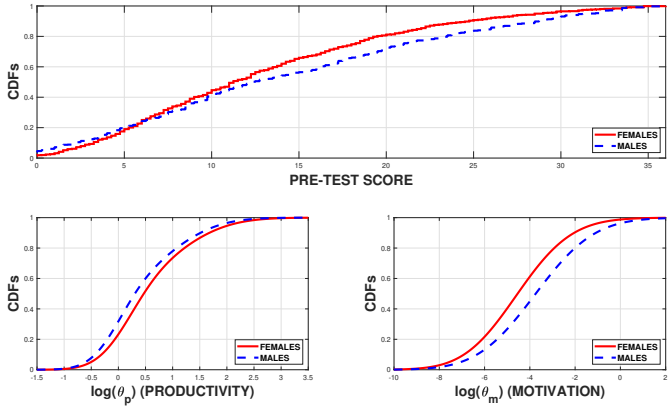
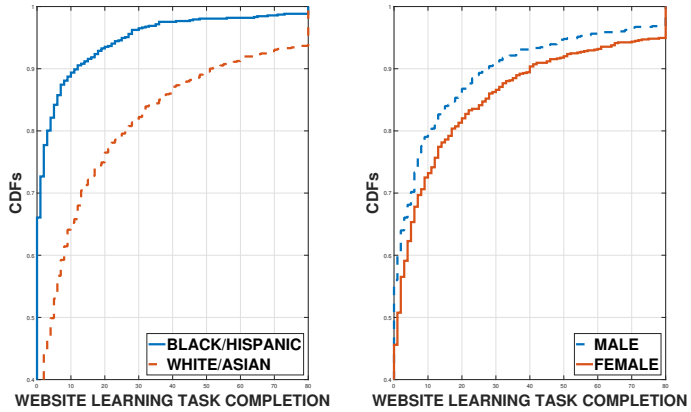


FIGURE 9. Website Task Completion by Gender and Race



A.3. **Additional Tables & Figures.**

TABLE 10. 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:	τ_0	τ_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 $c'(0) = 1)$.

FIGURE 10. Empirical Bayes Shrunk Type Estimates

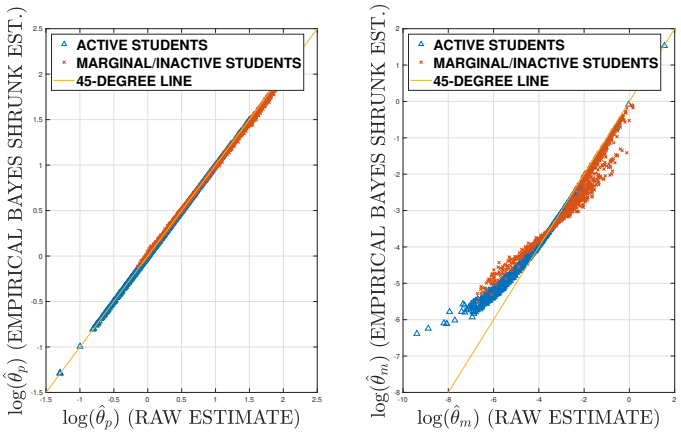
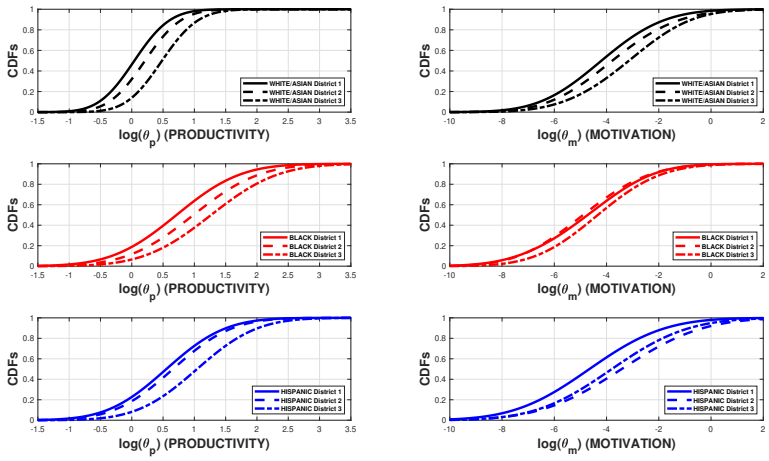


FIGURE 11



Appendix B. Online Supplement to accompany *Disentangling Motivation and Study Productivity as Drivers of Adolescent Human Capital Formation: Evidence from a Field Experiment and Structural Analysis*, by Christopher Cotton, Brent Hickman, John List, Joseph Price, and Sutanuka Roy

B.1. **Common Core Math Subject Sub-Categories.** We used standard Common Core subject definitions (accessible at `https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf` as of January 2023) to classify and organize pedagogical content on our website and proficiency assessments. These subject definitions are grade-specific, but with considerable overlap in the themes and 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.

B.2. **Internet Access Issues.** 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 11 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, 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,

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 document (accessible at `https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf` as of January 2023). 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*.

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.

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	-2.45	*(1.946)*	[-6.24,1.36]	-2.33	*(1.947)*	[-6.14,1.49]
District2	-11.16***	*(1.762)*	[-14.62,-7.71]	-8.50***	*(2.168)*	[-12.82,-4.32]
District3	-17.24***	*(2.506)*	[-22.16,-12.33]	-12.03***	*(3.588)*	[-19.07,-5.00]
nbhd_income	-3.9×10^{-5}**	(1.8×10^{-5})	$[-7.3,-0.4]\times10^{-5}$	2.4×10^{-5}	(7.6×10^{-5})	$[-1.1,1.5]\times10^{-4}$
$nbhd\_income^2$	—			-2.1×10^{-10}	(2.1×10^{-10})	$[-6.3,2.2]\times10^{-10}$
(hmwk_time/day)	-1.76	*(1.169)*	[-4.05,0.53]	1.99	*(2.842)*	[-3.58,7.57]
$(hmwk\_time/day)^2$	—			-2.11**	*(0.933)*	[-3.94,-0.28]
math_attitude	2.55***	*(0.391)*	[1.78,3.31]	1.40***	*(0.489)*	[0.44,2.36]
$math\_attitude^2$	—			0.38***	*(0.096)*	[0.19,0.56]
Contract2	2.49**	*(1.227)*	[0.08,4.89]	2.44**	*(1.220)*	[0.05,4.83]
Contract3	4.73***	*(1.226)*	[2.33,7.13]	4.71***	*(1.217)*	[2.32,7.10]
Constant	15.68***	*(3.326)*	[9.16,22.20]	8.81	*(5.654)*	[-2.27,19.89]
Gender/Race Dummies	NO			YES		
R^2	0.126			0.144		

Notes: We follow typical "star notation" for statistical significance; "***" denotes significance at the 1% level, "**" denotes significance at the 5% level, and "*" denotes significance at the 10% level.

B.3. **Structural Estimator Technical Details.**

B.3.1. *Dealing With Upper-Tail Mass Points of Learning Task Accomplishment.* We have a small mass of students who achieve full output $A_i = 80$ on the website, as can be seen in Figure 2. 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 $A_i > 80$ if given the chance.[1] 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 85^{th} 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

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

tables included 3 steps under contract 1 (with the smallest upper mass point), and 5 steps each for contracts 2 and 3. Figure OS.2 in the online supplement plots the extrapolated upper tails against the empirical CDFs of A. After discretizing the upper tail, for each 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 $\widetilde{G}_a$.

B.3.2. *Numerical Solution Approach.*

B.3.3. *Standard Errors.* For the empirical model of student time allocation and for the Tobit ML decomposition of student traits, we bootstrap all standard errors. Our block-bootstrap procedure is designed to mimic our randomized sampling procedure (discussed in Section 3.2.5) 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.[2] 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 single block-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.
(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 80 minutes each, including an adaptive multiple re-starts algorithm to ensure quality of the final solution. The Tobit ML estimator took a similar amount of time to converge for each bootstrap iterate. We chose 280 bootstrap samples for the time allocation model, and 280 bootstraps for the Tobit ML model.

[2]Due to a sparsity of Blacks and Hispanics in District 1 and a sparsity of Whites and Asians 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.

For standard errors on student fixed effects, we first bootstrap all common parameters. Then, we combine the bootstrapped parameter samples, $\left\{\tau_0^{(s)}, \tau_1^{(s)},\ \varphi^{(s)},\ \boldsymbol{\gamma}_c^{(s)}\right\}_{s=1}^{280}$, etc., with an individual's observables, $\left\{\{\tau_{a_i=1}^{A_i}\}, T_i, A_i, \boldsymbol{X}_{pi}, \boldsymbol{X}_{mi}\right\}$, to compute bootstrapped fixed effect estimates $\left\{\theta_{pi}^{(s)}, \theta_{mi}^{(s)}\right\}_{s=1}^{S}$. These within-student bootstrap samples of fixed effects are then used to compute standard errors, inverse variance weights, and EB shrinkage forecasts. We compute heteroskedasticity-consistent standard errors and hypothesis tests for production technology parameters in the usual way.

B.4. **Tobit Estimator Technical Details.** 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 (8), 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* $(\boldsymbol{\beta}_p^{HGl}, \boldsymbol{\beta}_m^{HGl}, \boldsymbol{\Sigma}^{HGl})$, $l=1,2,3$, by the following:

(A) $l = 1$: optimize (8) 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 (8) 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 (8) 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* $(\boldsymbol{\beta}_p^{SGl}, \boldsymbol{\beta}_m^{SGl}, \boldsymbol{\Sigma}^{SGl})$, $l = 1, \ldots, 6$, 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 (8) 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 $(\boldsymbol{\beta}_p^{**}, \boldsymbol{\beta}_m^{**}, \boldsymbol{\Sigma}^{**}, \mathcal{L}^{**}) = \left\{ (\boldsymbol{\beta}_p^{rGl*}, \boldsymbol{\beta}_m^{rGl*}, \boldsymbol{\Sigma}^{rGl*}, \mathcal{L}^{rGl*}) : \mathcal{L}^{rGl*} = \min_{r'=H,S;G'=I,\ldots,VII;l'} \{\mathcal{L}^{r'G'l'*}\} \right\}$.

(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 (8) is smooth enough for derivative-based quasi-Newton methods to function properly (see footnote 31), 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 particularly useful when local optima exist within the vicinity of a 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.

As a final practical note, if the above process is executed exactly as described it can be thought of as a Gauss-Seidel approach to solving for a fixed point corresponding to a global optimum. However, 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 across values of $r = H, S$; $G = I, \ldots, VII$; and l. This adjustment to the process would bring it more in line with a Gauss-Jacobi fixed point approach.

TABLE OS.3. 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%	2%	4%	24%
% Meeting/Exceeding Expectations on State Standardized 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 2 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.

TABLE OS.4. 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.1. Conditionally Heteroskedastic Work-Time Shocks

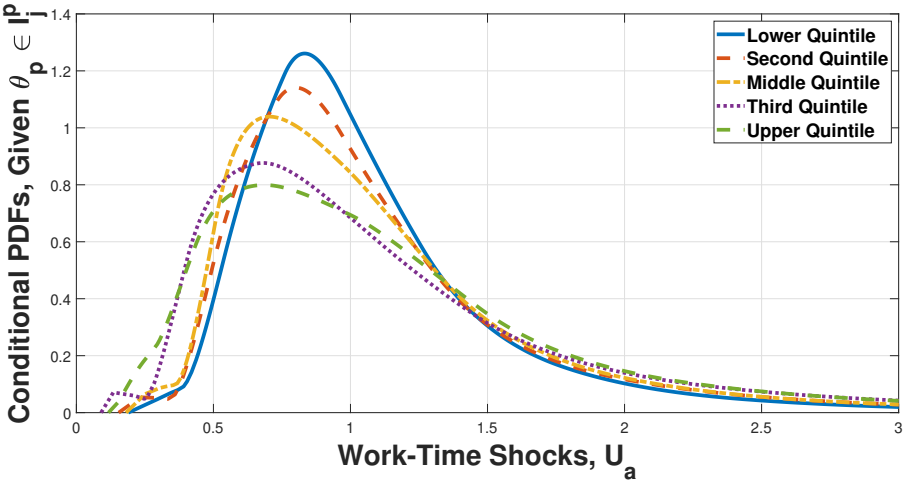


FIGURE OS.2. CDF Smoothing and Upper Tail Extrapolation

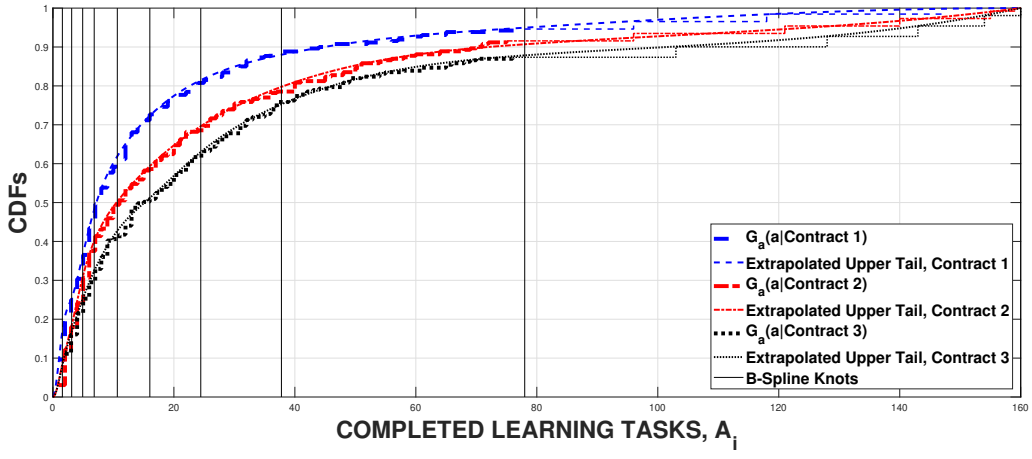
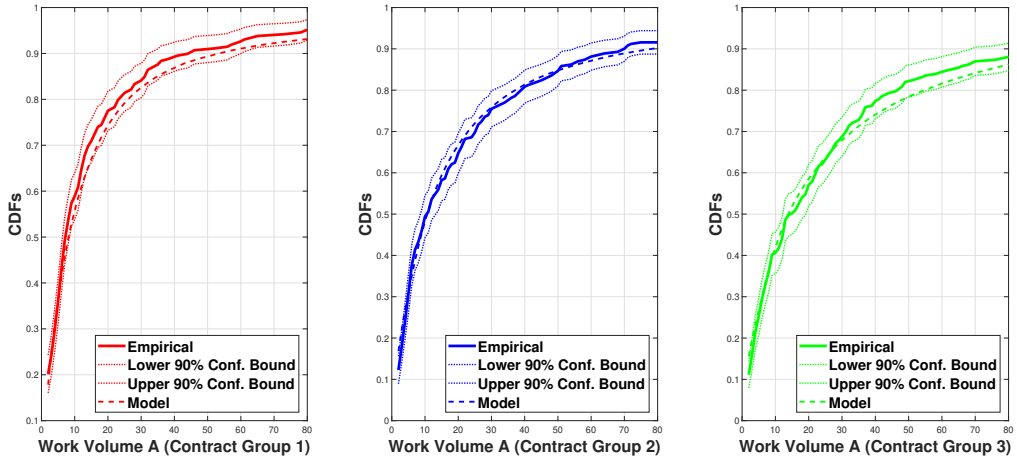


FIGURE OS.3. Cost Model Fit



B.5. **Supplemental Tables and Figures.**