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

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

Abstract. We conduct a field experiment across three diverse school districts to structurally identify student motivation and study productivity parameters in a model of adolescent human capital development. By observing study time, homework task completion, and test results, we can identified individual and demographic variations in motivation and study time effectiveness. Struggling students typically do not lack motivation but rather struggle to convert study time into completed assignments and proficiency improvements. The study also attending a higher-performing school is associated with both higher productivity and higher motivation relative to peers with similar observables in lower-performing schools. Counterfactual analyses estimates that school quality differences account for a substantial share of the racial differences in test scores, and considers the impact of alternative policies aimed at reducing racial performance gaps.

Keywords: Human capital; field experiment; structural econometrics; psychology of education; student motivation; school districts

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

date: **July 2023**. **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 Nicholas Buchholz, Greg Sun, Barton Hamilton, Caroline Hoxby, Chris Taber, Jeffrey Smith, Samuel Purdy, Mary Mooney, Felix Tintelnot, Aloysius Siow, Angela Duckworth, and Joseph L Mullins for particularly helpful conversations, as well as 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 for 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, Andrew "Rusty" Simon, and Diana Smith. Finally, we are indebted to an army of undergraduate 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.

Cotton: Department of Economics, Queen's University, Kingston, Canada, cc159@queensu.ca.

[†]Hickman (corresponding author): Olin Business School, Washington University in St. Louis, hickmanbr@wustl.edu.

List: Department of Economics, University of Chicago and NBER, jlist@uchicago.edu.

Price: Department of Economics, Brigham Young University, Provo, UT, joseph_price@byu.edu.

Roy: Australian National University, sutanuka.roy5@googlemail.com.

1. Introduction

Suppose that Anthony, a 6th-grade student, does not regularly complete his mathematics homework and performs poorly on exams. Seeing this, one may think that Anthony lacks the motivation to put the time and effort into studying math, not perceiving enough value in completing assignments, performing well on tests, or learning the material relative to the opportunity costs of his time. If this is the case, it may be possible to increase Anthony's academic performance by increasing his motivation by providing incentives or information (e.g., about the future returns from schooling) to him or his parents. However, a lack of motivation is not the only possible explanation for Anthony's observed choices and outcomes. Anthony might be more than willing to spend time studying math, but he might struggle to convert his study effort into completed homework or increased learning. If he lacks foundational math or literacy skills, adequate study support, high-quality instruction in school, or accommodations for a learning disability, then spending even large amounts of time on math may not be enough for Anthony to finish assignments and raise his grades. If so, Anthony may rationally decide not to bother with the assignment. In this case, a very different intervention would be appropriate, and merely providing additional incentives or information is unlikely to substantially improve Anthony's academic outcomes.

The challenge for policymakers, educators, and researchers comes from an inability to directly observe whether low-performing students lack motivation—i.e., willingness to allocate a fixed quantity of time to study—or whether they struggle with low productivity—i.e., the rate at which they convert their time inputs into measurable academic success. Because typical observational datasets include only coarse measures of student time inputs (if any at all) and tend to focus more on output measures such as grades/exam scores, both explanations—i.e., low motivation and low academic productivity—are either observationally equivalent or difficult to separately identify.[1] Moreover, there is no data set we are aware of that would allow for estimating a direct mapping between day-to-day learning activity (or lack of it) and incremental skill gains. This inferential problem is a formidable barrier to understanding student challenges (both at an individual and group level) and designing effective education policy.

To overcome these challenges, we execute a structural empirical study using a field experiment involving 1,676 adolescent math students in partnership with their schools and teachers.[2] The setting of a natural field experiment allows us to observe how study effort responds to incentives, how effective students are at converting time inputs into completed assignments, and how additional study time leads to test score gains. Structural econometric methods allow us to directly quantify individual labor-supply elasticity and study-time productivity measures for each 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 study-time productivity differ by initial test performance, demographics, socio-economics, school quality, and other factors, which facilitate counterfactual analyses of policies aimed at reducing education performance gaps. Our field experiment produces a uniquely-rich set of student-level variables and varying incentives that are well-tailored to solve various empirical confounds present in typical observational data from educational

[1] Perhaps the most closely related papers to ours are Agostinelli and Wiswall (2022), which uses NLSY data with observations at two-year intervals, and Del Boca, Flinn, Verriest, and Wiswall (2019), which uses PSID data including child time-use variables. Unfortunately, neither of these datasets (or any others we know of) includes time-use data and exogenous incentive variation, the two features 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.

settings. Our field experiment is designed to generate the individual level observable data required to estimate a novel quantitative framework of day-to-day learner labor-supply (e.g., study-time) decisions.

For analysis of our uniquely rich field experimental data, we develop a novel quantitative framework for studying day-to-day learner labor-supply choice. Taking cues from the psychology literature (e.g., Carroll (1963); Eccles and Wigfield (2002)), we model each student as having two idiosyncratic characteristics that govern productivity (i.e., rate of progress through learning assignments) and the opportunity cost of time. A child's utility costs of spending time on math are convex, meaning that she becomes increasingly less willing to continue work as her total time allotment to math study grows. This model feature admits various natural interpretations, including physical/mental exhaustion, or marginal value of non-math time rising as other activities become increasingly crowded out of the child's daily schedule.

A central theme of this framework is that the 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 time inputs. If Anthony and Joseph both earn an "A" grade in a math course then, all else equal, they both see the same improvement in their chances of landing a particular job, gaining admission to college, etc., even if achieving the "A" grade took Joseph only 100 hours of work, while it took Anthony 500 hours. Within our model, while a child's motivation characteristic alone determines his willingness to spend an hour studying math, under piece-rate incentives his productivity characteristic will also play a central role in his decision of whether to spend *enough time* on math to achieve at a high level. Our model shows how a child with high motivation (i.e., relatively low cost of spending an extra hour on math) may still make choices such that he will appear to be academically withdrawn, 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 will be informative for understanding the contribution of our framework and methods. A branch of this literature focuses on parental investment in child human capital and/or policy interventions like financial resources or incentives for parents; e.g., Cunha, Heckman, and Schennach (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 this final missing link is uniquely crucial among the three components, as learner choices are the one truly indispensable input. For example, some of the harm from low school quality may be offset by intensive parental inputs. Moreover, many public education policies, such as universal pre- K, free and reduced-price lunch, and after-school programs, are geared toward (partially) offsetting scarce or lacking parental inputs in a child's education. If parental and school educational inputs are lacking, personal effort on the part of the learner may even substitute for both. Prominent historical examples include Alexander Hamilton, Frederick Douglass, and Abraham Lincoln, while contemporary examples abound, including authors Tara Westover and Jeanette Walls. On the other hand, if the learner himself/herself is unwilling or incapacitated from contributing to the parent-child-school partnership, it is difficult to imagine a viable compensating factor in human capital production. Our research is designed to provide a novel and detailed window into the learner's rational decision process of *self-investment.*

A newer branch of the literature focuses on experimental studies of children's responsiveness to incentives as a means for spurring academic improvement, including Fryer (2011), Fryer (2016) Levitt, List, and Sadoff (2016), Burgess, Metcalfe, and Sadoff (2016), and Cotton, Hickman, and Price (2022). While compelling as a growing body of evidence, Fryer (2011) articulated well the main 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 Cotton et al. (2022), this literature is not designed to 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 into the underlying factors that affect student study choices, rather than studying the impact of a specific incentive scheme or policy. In particular, 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 cross-student variation is attributable to observable external differences (e.g., family background, outside options for time-use, 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 that dominate academic incentives?

We find strong empirical evidence that a large fraction of struggling students are not less motivated than their higher performing peers. Rather, their struggles stem from difficulty in converting effort into success. We also document surprisingly large productivity heterogeneity among students whose observable achievement is well above average, including many with study productivity well *below* average. Conversely, we also find nearly equal heterogeneity among students whose observed achievement choices are well below average, including many that are quite productive in terms of converting time into completed learning tasks. 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 marginal increase of additional math activity from 15 minutes per day to 30 minutes per day (i.e., a doubling of the child's marginal time commitment) causes monetized utility costs to rise by a factor of more than three.

In a series of secondary analyses, we investigate to what extent a child's observable life circumstances can explain productivity and motivation heterogeneity. While a large fraction of this heterogenetiy appears truly idiosyncratic to the child and context-independent—53% for productivity and 64% 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. One of our main empirical findings is that school quality is a substantial factor for augmenting both academic productivity (i.e., rate of progress through learning tasks) and motivation (i.e., willingness to spend a fixed quantity of time on math study).

In another set of secondary analyses based on our structural estimates, we investigate the mapping between student productivity/motivation traits, interim learning activities, environmental factors, and measurable incremental gains in math skill. These are based on pre-test scores and post-test scores logged just before and just after our sample period, with opportunities for incentivized learning activity in the interim. We find that volume of successfully completed practice problems is the primary driver of incremental skill gains, with time spent actually serving as a (slightly negative) mitigating factor. Moreover, student productivity not only spurs higher rational choices of learning task accomplishment, but also it fundamentally alters the shape of the human capital production technology as well. More productive students not only progress through math practice problems at a higher rate, but they are also able to more effectively convert a fixed volume of math practice problems into permanent skill gains.

The empirical analysis also points toward a decreasing returns to scale learning technology, whereby math score gains of a fixed margin become harder to accomplish as a child's initial proficiency level rises. Once again, we also find strong evidence that school quality plays a significant role in shaping human capital production technology. Not only does it advantageously influence a child's productivity and motivation, which in turn spurs greater volumes of math study choices, but also, it increases the rate at which a fixed quantity of math practice problems being solved are converted into permanent math skill gains. Given geographic racial sorting patterns common to Chicago and most other US cities, a large fraction of the racial achievement gaps in our sample population can thus be causally tied to school quality differentials.

Finally, in a series of counterfactual model simulations we run the thought experiment of holding all aspects of Black/Hispanic students' lives fixed—from family background and affluence, to attitudes and consumption/time-use opportunities, to existing external academic incentives—while replacing their observed school of residence with one drawn at random from the distribution of school assignments enjoyed by their White and Asian peers. We then run the school quality differential through various aspects of the model and re-compute how the child's productivity, choices, and outcomes would change. The most striking aspect of this exercise is how much more responsive these children become to external enticements toward study. This result empirically highlights the implications of education systems based on piece-rate incentives: when a child's rate of progress through learning tasks significantly improves, the salience of his existing academic incentives drastically increases.

Our study highlights how important the learning environment is in shaping both rational academic choices and performance. The school district where one enrolls matters more for one's test scores than race. Even when controlling for a wealth of individual and family characteristics, we find that those who attend school in more-affluent districts have higher test scores and are better able to convert study effort into score improvements. Even highly-motivated and productive students tend to perform significantly better when they attend a high-performing district than a low-performing district. On the other hand, the decreasing returns to scale production technology implies that the overall social value of improving school quality is highest among struggling students who lack educational resources to begin with.

This paper is structured as follows. Section 2 outlines the quantitative 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 a host of observable environmental factors and student characteristics. Section 6 illustrates how the inferred student motivation and productivity data allows for a rich set of analyses, exploring how these and other factors relate to initial mathematics proficiency, the interaction between study effort and academic progress, and a series of counterfactual analyses highlighting the role that school quality plays in driving racial performance gaps and comparing the impact of policies that aim to reduce these gaps. Section 7 concludes. An included appendix contains additional technical details, graphs, and tables and on online Appendix contains additional technical details of our research design and methodology.

2. A Theoretical Framework for Learner Self-Investment

We develop a quantitative model of study effort and academic outcomes that takes cues from the qualitative frameworks of "mastery theory" (e.g., Carroll, 1963) and "expectancy-value theory" in education and psychology (e.g., Eccles et al., 1983; Wigfield, 1994; Eccles & Wigfield, 2002; Wang & Degol, 2013). In these literatures, the process of skill acquisition is viewed as a sequence of learning tasks: a child learning algebra attends class each day, is assigned a number of practice problems for homework, chooses how much of each assignment to complete, and then iterates the process anew each day until the algebra course concludes. Regardless of raw talent, anyone wishing to acquire a new stock of skills or knowledge must traverse some critical mass of learning tasks, or they will not develop competency in algebra. Thus, human capital attainment ultimately hinges on a series of high-frequency, but individually low-stakes, 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. The core model primitives include a 2-dimensional vector of child characteristics, which shape various aspects of learning, including the mapping between effort and rewards. While the core of our empirical exercise (and the sole focus of the first half of the paper) is unobserved child heterogeneity and its role in decision making, Sections 5 and 6 show how the structural primitives of the model we develop here 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 Student Study Choices.

We assume 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 (Carroll (1963); Wigfield and Eccles (2000)). We refer to these components of incentives as *study-time productivity* and *motivation*, which we now formally define. For each child, indexed by i, let $A_i \in \mathbb{N}$ denote the total number of learning tasks that i completes within a fixed period of time. The precise definition of a "time period" is not crucial, provided that it is a short enough span so that a child's decision-relevant characteristics can be thought of as fixed and stable (at least to a first-order approximation) within a period. Once the length of a period is specified—e.g., a day, a week, a month, or perhaps a semester—our model is to be interpreted as one of within-period decisions, and the costs within the model can be thought of as implicitly representing opportunity cost of allocating a certain fraction of within-period time to study, while necessarily crowding out other activities, given a finite endowment of time.

Each individual task, chronologically indexed by $a_i = 1, 2, \ldots, A_i$ for child i, represents some discrete unit of work, which can be specified as finely as a single math problem, or as coarsely as an entire assignment or block of assignments. Completion of these learning activities builds skill proficiency, which is of ultimate interest to policy-makers. We study the mapping between work and proficiency gains in Section 6, but at present we begin by focusing 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 the parameter θ_p, but for ease of discussion we henceforth refer to it simply as "productivity".[3] 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:

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

Here, $\tau_i(a;\theta_{pi})$ is the time required for student i to complete her a^{th} learning task. The τ_0 term is mean completion time across all students, τ_1 is a "startup cost," which applies only to the first task of the period, and the productivity fixed effect θ_{pi} scales this mean 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 $\varphi < 0$) as in-period task volume grows. Thus, the experience curve term allows for short-run gains (or losses) in study productivity, despite θ_{pi} being fixed within-period.[4]

The transitory study-time 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 learner approaches a given problem with dread, only to be pleasantly surprised at a quick turnaround time; other times a learner's initial optimism melts away as a given problem unexpectedly drags on. By convention, we use upper-case U to denote the random variable and lower-case u for specific realizations; we also use subscripted "F" to denote exogenous distributions, while subscripted "G" indicates a distribution of some endogenous object. We assume the distribution of the shock 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 support $[\underline{u}, \overline{u}] \subset \mathbb{R}_+$.*

A child's time may be put to various alternate uses, so math study is costly. For the purpose of our field experiment and empirical model, we focus specifically on math study; however, the model admits various interpretations about what A_i and T_i represent. The could alternatively represent general schoolwork, in which case the outside option for time encompases all non-scholastic activity (e.g., sleep, chores, relaxation, socializing, recreation, etc.). On the other hand, 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 say, Science homework, and diverting time toward math makes it more difficult to attain rewards for Science achievement or to avoid punishment for lack thereof. Thus, we empirically formalize how costly it is to individual students to substitute time away from the most profitable outside use (including all other academic activity like homework in English, Science, etc.) and toward math learning.

[3] 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.

[4] We find very modest experience-curve effects: $\hat{\varphi}$ implies a 5% reduction after doubling output, or $\frac{\tau_i(2a_i;\theta_{pi})}{\tau_i(a_i;\theta_{pi})} \approx 0.95$.

Definition 2.2. *(Inverse) Motivation*, denoted $\theta_{mi} > 0$, indexes idiosyncratic labor-supply costs, and reflects student 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 "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: $C_i(T_i; \theta_{mi}) \equiv \theta_{mi} c\,(T_i)$. Assumption 2 establishes regularity conditions on costs 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 impose scale and location normalizations of $c(0) = 0$ 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 by math, 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 holistic 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 holistic 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} in a more "reduced-form" way.[7] In Section 5, we draw upon a wealth of student-level observables to empirically tease apart the various factors that contribute to student types, and decompose $(\theta_{pi}, \theta_{mi})$ into predictable and idiosyncratic components. However, for the purpose of the formal structural model, θ_{mi} can 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 could be made about 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. *For each child i, a piece-rate payoff function $\Pi_i(A)$ governs external incentives. Payoffs are increasing in A and bounded: there exists $\overline{\pi} < \infty$ such that the difference $\Pi_i(A) - \Pi_i(A-1) \leq \overline{\pi}$, $\forall A \geq 2$.*

[5] Unbounded marginal costs ensures finite optimum study choices under any finite reward scheme. For example, if a period is interpreted as a week, and t is measured in hour units, then one might naturally assume $c'(t) \to \infty$ as $t \to 168$ as an Inada condition on utility. This limiting choice would entail 7 full days of completely 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 human capital.

[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 incurs larger psychic costs of exerting effort solving math problems, relative to j. Alternatively, j's home/school environment may engender norms that shape perceptions of work as being less onerous than i's perceptions.

The bounded incremental payoffs assumption is needed to ensure a well-posed student decision process. Intuitively, piece-rate incentive schemes may encompass all external "carrots" and "sticks" presented to child i by her home, school, and community. For example, parents may inculcate in her a positive intrinsic valuation of achievement, they may offer tangible monetary or non-monetary rewards, or they may stipulate punishments that can only be escaped through regular completion of work and/or achievement benchmarks. A child's school rewards her for homework completion with grades, and may further motivate her regular coursework via pre-announced exams. These cumulative grades and test scores will in turn determine her future education and career prospects. Additionally, businesses, organizations, and colleges may offer merit-based admissions, internships, or scholarships that represent expected flows of future monetary and psychic income from a desirable career (Becker (1993)). Conversely, failure to perform academically may be discouraged by the prospects of an undesirable or less stable career path.

In choosing total work time T_i and task completion A_i, a student solves an optimal stopping problem by recursively comparing benefits and costs of finishing an additional learning task while accounting for randomness in completion times.[8] Given $(a_i - 1)$ completed learning tasks, a student optimally determines the maximum time, $t^*_{a_i}$, that she is willing to devote to completing the a_i^{th} task. Due to random study-time shocks U_{a_i}, equation (1) implies success probability on task a_i, given t units of time input, is $F_s(t; a_i, \theta_{pi}) \equiv \Pr\big[\tau_i(a_i;\theta_{pi}) \le t | a_i\big] = \Pr\big[U_{a_i} \le \frac{t}{\theta_{pi}\tau_0\tau_1^{\mathbb{1}(a_i=1)} a_i^{-\varphi}}\big] = F_u\Big(\frac{t}{\theta_{pi}\tau_0\tau_1^{\mathbb{1}(a_i=1)} a_i^{-\varphi}}\Big|\theta_{pi}\Big)$, with first derivative denoted by $f_s(t;a_i,\theta_{pi})$. Thus, a learner's decision problem is defined by the Bellman equation,

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

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

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

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.

From this it follows that her optimal choice t^*_a is implicitly defined by the first-order condition

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

Intuitively, if she achieves the a_i^{th} success with some work 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 $t < t^*_{a_i}$ units of time does not reach the completion state on task a_i, she continues to work and equation (4) says that her stopping time $t^*_{a_i}$ dictates the point at which the expected marginal benefit of continuing work (including retention of future payoff opportunities) is exactly offset by the 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

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

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 completion-time shocks encountered along the way to her stopping point. For notational compactness we suppress this dependence, but note that the important implication is that, holding fixed a given $(\theta_{pi}, \theta_{mi}, \Pi_i)$ triple, there will be a non-degenerate distribution of final within-period choices, whose joint distribution is denoted by $G_{ta}(T_i, A_i|\theta_{pi}, \theta_{mi}, \Pi_i)$.

Two key model predictions will be relevant to identification of the structural model later on. First, if a child continues on to the $(a_i+1)^{\text{th}}$ task, note an important shift in her decision problem: on the previous task 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 elevated marginal costs of continuing math work, 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}): reductions in mean completion times lead to first-order dominance shifts, or $\theta_p < \theta'_p \Rightarrow G_a(a|\theta_p, \theta_m) \leq G_a(a|\theta'_p, \theta_m)$, $a \in \mathbb{N}$.

2.2. **The Significance of Piece-Rate Incentives.**

Our choice to represent work compensation as piece-rate—that is, $\Pi_i(\cdot)$ is a function of completed tasks, A_i, rather than time inputs, T_i—was deliberate as these schemes are 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, it does not matter for rewards if this result required a total of 40 hours study for *Jane*, but only 20 hours for *Tabby*. Both children would receive the same grade as a result of their identical performance record and, ceteris paribus, that grade will map equally into increased likelihood of obtaining their desired college seat, scholarship, internship, job, etc. This modelling choice is not only empirically relevant, but also it has profound implications for how incentives interact with a child's unobserved, idiosyncratic characteristics. 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) where children are rewarded for outputs rather than inputs, they recognize that rewards not only require time investment, but successful conversion of time into final results as well. Thus, 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 decisions immediately calls into question prevailing wisdom behind labels that are often applied to different students, based on observed behaviors. For example, if we see that *Jane* turns in only 50% of her assignments, while *Tabby* completes all of them, many practitioners and researchers would simply conclude that *Jane* is "unmotivated" for math studies, while *Tabby* appears "well-motivated." It is certainly plausible that the observed behavior difference could stem from Jane being less motivated $(\theta_{m,Tabby} < \theta_{m,Jane})$. However, under the model it is also plausible that Jane turns in less homework because she is sufficiently less productive $(\theta_{p,Tabby} < \theta_{p,Jane})$, even if she is *more* motivated than

Tabby ($\theta_{m,Tabby} \geq \theta_{m,Jane}$). The model shows that both of these explanations are observationally equivalent given the single raw data point of *Tabby*'s and *Jane*'s study choices or academic performance.

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 observable output, and in that sense 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, costs incurred, etc.) and observable output (e.g., grades, homework, etc.). This is a problem for educators and policy-makers alike: the two opposing explanations for *Tabby*'s and *Jane*'s choices would imply entirely different interventions to help children like *Jane* succeed academically.

2.3. **Discussion on Modelling Choices.** Before moving on, it will be useful to briefly discuss some aspects of our modelling approach. First, recall that our goal is to study short-run, high-frequency, adolescent decision processes, so there are no formal "future periods" in the model, aside from the chronological indexing, "a," of learning tasks worked on. 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 constitutes a non-stationary dynamic program because of cost convexity combined with with history-dependence of the state variables $(a_i-1, T_i(a_i-1))$ for the current unit of work a_i. The continuation value argmax (equation (4)) updates continuously as time accrues working on task a_i, and the final distribution of payoffs on future units of work $\widetilde{a} > a_i$ is not known until the moment when the student finally completes the current task.[9] Although the model is computationally taxing for the researcher, there are several advantages to modeling short-term leisure-study decisions this way, the most important being that it requires only simplistic thinking on the part of the adolescent decision-maker. The optimal stopping model merely assumes that the child apprehends three basic pieces of information at each point in time: (*i*) her marginal incentives to complete another task, (*ii*) the unpredictable variability of completion times for current and future tasks, and (*iii*) how taxed she feels from previous work leading up to the present. In short, a child solving for optimal (T_i, A_i) need only be aware of her current feeling and of her ability to continue productively for a little bit more work.

An alternate model where children make a one-shot decision on achievement target A_i at the beginning of the period would be computationally simpler, but would impose much stronger assumptions on decision-making. It would require that adolescents 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. If a child completes A_i tasks in an unexpectedly short amount of time, she would have incentive to work beyond the original target A_i: her marginal cost of time $c'(T_i(A_i))$ would still be relatively low, allowing her to reap high expected marginal payoffs. By similar logic, a child encountering a particularly bad string of shocks along the path to her target A_i would wish 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.

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

3. A Field Experiment to Identify Student Motivation and Productivity

Observational equivalence between drastically opposing explanations for student behavior means that the model of adolescent study-leisure choice with two-dimensional heterogeneity is *not* identified from standard observational data. This fact motivates our field-experimental design, carefully crafted to identify multiple dimensions of student observables and exogenous variation that are hard to come by or non-existent in passively collected data. Our research design forms part of a nascent literature that employs experimental methods for the relatively novel purpose of 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 (2019), Bodoh-Creed, Hickman, List, Muir, and Sun (2023)).[10]

Our experiment was *not* designed to test the impact of paying students a certain amount of money to study (e.g., as in Levitt, List, and Sadoff (2016)). Rather, our purpose is to empirically model underlying mechanisms behind a student's daily effort choices under *any* incentives. Our strategy uses the student choice model as a basis for an econometric framework, and experimental methods shape a data-generating process with the 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 some common experimental designs. For example, we need not specify a control group to serve as an empirical baseline. Rather, we simply specify multiple treatment groups that are exposed to exogenously differing levels of incentives in order to identify idiosyncratic elasticities and trace out the curvature of the labor-supply cost function. In that sense, our experimental variation is similar in spirit to *A/B testing* methods commonly used in marketing and user-experience optimization in e-commerce.

Our approach to quantifying unobserved student traits combines standard panel-data methods with recent econometric theory developed for using discrete instruments to identify continuously varying unobserved heterogeneity (Torgovitsky (2015) and D'Haultfoeuille and Février (2015, 2020)). Our field experiment also allows us to craft a data-generating process to be as true to a student's everyday academic environment, choices, and experiences as possible. We fully discuss our experimental design in Section 3.2, but first it is useful to explore conceptually the identification strategy that guides it.

3.1. **Identification Strategy.**

For intuition on how we quantify unobserved student traits, consider a hypothetical "ideal" experiment involving students, $Tabby$ and $Jane$, who perform poorly on a standardized math exam. The exam score indicates that each student is struggling, but it offers no insights as to why. To answer this question, the researcher obtains two clones, call them $Tabby^*$ and $Jane^*$—i.e., identical in biology, ability, preferences, family background, etc.—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 mathematics textbook. There is also a couch with a web-enabled smart-TV, a video gaming system, a smart phone connected to social media, and other leisure opportunities. 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

[10] Augenblick et al. (2015) used a framed field experiment to point-identify dynamically inconsistent time preferences of college students. Rao (2019) executed a framed field experiment to identify demand curves for discrimination among two groups with differing exposure to social diversity. DellaVigna et al. (2022) and Hedblom et al. (2019) use natural field experiments to identify individual-level unobserved worker characteristics. Bodoh-Creed et al. (2023) propose econometric methods to identify robust bounds on consumer demand counterfactuals under out-of-sample prices, using field experimental data.

defined by some concrete criterion). The researcher explains that the children are free to allocate their time in any way, working through as many or as few math exercises as 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 a sequence of 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 heterogeneous effective mean hourly wage rates; e.g., suppose $Tabby$ works fast enough to earn \$15/hour on average, whereas $Jane$ works slower and can garner only \$12/hour on average. Note that all differences in mean hourly wage between $Jane$ and $Jane^*$ are due solely to their piece-rate offers $\pi < \pi^*$, since they are otherwise identical and have the same productivity $\theta_{p,Jane}$. Since $Jane$ ($Jane^*$) produced more output than her same-piece-rate counterpart $Tabby$ ($Tabby^*$) despite lower hourly compensation, it must be that $Jane$ is more willing to allocate time toward math learning than $Tabby$ (i.e., $\theta_{m,Jane} < \theta_{m,Tabby}$).

The piece-rate shift from π to π^* can identify 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 shape of $c(t)$, independent of $Tabby$'s and $Jane$'s idiosyncratic traits. E.g., $Tabby$'s output increased by 40% while $Jane$'s output under the same proportional wage increase rose by only 30%, indicating that marginal costs must be higher from $Jane$'s baseline output of 10 assignments, relative to $Tabby$'s baseline of 5 assignments. Inferences about the form of the common cost schedule become richer as the experiment is repeated with an large set of $Tabby$'s and $Jane$'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 leisure preference 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.

While informative as a thought exercise, much is obviously infeasible or unethical about the above "ideal" experiment. However, with field experimental methods and web-based technologies, one can capture the essential elements while maintaining a level of realism and familiarity that would be impossible within a controlled laboratory setting. One can easily "clone" groups of students through individual-level randomization. While no two groups will contain identical copies of the same child, the overall 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., classroom instruction and graded homework assignments), a question arises: how do we interpret experimentally observed (extracurricular) math activity, given that concurrent, formal coursework and its outside incentives (by parents, schools, communities, etc.) are unknown to the researcher? A major challenge to any empirical modelling in this context (including our web-based experimental approach) is that the payoff function in its entirety, encompassing all external "carrots" and "sticks," is notoriously difficult to quantify, due to data limitations. Thus, our field experimental design solves this problem by

[11] As a side benefit, our web-based research design provides a proof of concept for powerful new diagnostic tools cheaply available to educators at scale, given recent dramatic increases in K-12 educational materials being housed online.

placing many different children on the same footing with a known external incentive function dictated by the research team. Moreover, structural identification is still possible, provided that the distribution of formal coursework activity is uncorrelated with incentives. Thus, individual randomization is crucial to ensure that experimental incentives are independent of a child's teacher, school, and unobserved external incentives provided by parents, schools, or 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 will still be informative for policy analyses focused on improving academic outcomes *relative to the status quo*. On the other hand, note that the interpretation of experimentally inferred productivity, θ_p, hinges only on the similarity between extracurricular incentivized math tasks and formal coursework.

Despite the caveat mentioned above, sufficiently 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 two-dimensional fixed effect encompassing all life circumstances—including default formal coursework commitment among other external and internal factors—relevant to student 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, preferences, family background, and other factors. Moreover, experimental estimates of unobserved student traits (and observed extracurricular math activity) may be formally projected onto standardized exam scores to produce policy-relevant comparisons among observably different students, and 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 also find evidence in our survey data consistent with their reports (see Section 3.3 below). Finally, our intuitive discussion above also glosses over an important issue of sample selection: what if $Tabby$ spent no time on math under incentive π, while her alter-ego $Tabby^*$ did some math time under incentive π^*? Holding (finite) 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. If variation in unobserved types is sufficiently high, this problem may persist even under generous piece-rate incentives. Some group of children may feel they are too inefficient or too averse to extra work (or both) to respond with positive labor supply. This is especially true if formal coursework already competes for Their time. 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 5th and 6th 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, including parent/child assent forms, proficiency exams, surveys, personalized incentives, etc. 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 such requests.

3.2.1. *Study Sample.* We partnered with three public school districts in the greater Chicago area during the 2013-2014 academic year. A total of 1,676 5thand 6thgrade students participated in the experiment, with 46% from *District 1*, and 27% each from *District 2* and *District 3*. These differed widely by population and administrative characteristics, which we summarize 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, District 1 was above-average on faculty compensation, teacher qualifications, fraction of budget spent on instruction, and student performance. District 1 was also well above the rest of the state in terms of financial resources per student. District 2 was remarkably close to the state averages on these dimensions, while District 3 lagged considerably behind on most, including the fraction of budget spent on instruction and student academic performance. District 3 had a relatively high operating budget per-pupil though, like many districts serving less affluent communities, it receives additional state funding for factors such as social workers, guidance counselors, ESL programming, lunch subsidies, non-instructional support programs, etc.

The populations these three districts serve are similarly ordered in terms of socioeconomic traits. District 1's student population is substantially more affluent by income and wealth, with all but 15% of its operating budget is derived from local property taxes. District 2 is closest to the state means, while District 3 is much less affluent by income and wealth, and has a relatively large fraction of students with limited English language proficiency (many from Hispanic immigrant families). Finally, the other striking difference is the racial sorting of the communities each district serves (see Table 8), which is typical of most major

[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 by signing and returning it, or by responding via email. On the first day of the study, students received an additional child consent form with similar information stated in age-appropriate language. This form emphasized that participation was optional and would not affect their academic standing; it 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 opted out or was opted out of the study.

metropolitain areas in the United states. District 2 has a racially diverse student body, while District 1 has mostly White and Asian students, and District 3 is mostly comprised of Black and Hispanic students.

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. Throughout the process, our web server monitored students' use of the site and tallied successful completions. We measured math proficiency using in-classroom assessments before and after the website was made available. Given our focus on structural identification of the student choice model primitives in this section, we defer discussion on exams and student survey data to Section 5, Section 6, and Appendix A. Teachers were the primary point of contact for student participants, with the exception of a technical support email account managed 24/7 by the research team.

3.2.3. *Mathematics Pedagogical Materials.* Proficiency assessments and website content were comprised of professionally developed, age-appropriate math materials. Specifically, 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.[13] We broke these materials up into individual math problems, resulting in 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*.[14] 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 both 5thand 6thgraders to the same materials. This pooling served multiple purposes. First, it provided a wide swathe of content well-suited to studying a diverse Chicagoland student population with considerable pre-existing proficiency heterogeneity. The goal was to provide a mix of both challenging and basic material. Second, it gave us a larger pool of learning materials from which to draw. Third, it facilitated an even 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 significantly by the striking similarities in curricula and common-core sample problems across grades 5 and 6: most 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 of Common Core recommendations for mathematics focus areas by grade, and present a side-by-side harmonization of the 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 and pedagogical materials sufficient to attract non-trivial participation from all segments of our sample population? If not, then poor experimental design would

[13] These 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 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).

[14] Common Core subject definitions for 5thand 6thgrades (`https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf` accessible as of January 2023) differ slightly; our 5-subject classification is a merging of the two (see Appendix B.1).

be manifest in the form of low statistical power within descriptive analyses and structural estimates. To the contrary, in Section 3.3 (especially Figure 1), Section 4.3, and Section 6 (especially Table 4) the data demonstrate that there were no undue scientific drawbacks to cohort pooling.

3.2.4. *Website Structure.* Our website was accessible through a login credential assigned to each student. The web server automatically tracked and recorded site activity for each child without affecting user experience in any perceivable way.[15] The website provided 80 learning tasks, each consisting of 6 multiple-choice questions from our bank of math problems. Six problems per task was chosen based on feedback from adolescent pilot-study subjects under similar incentives as in the main study. 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 was 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, and that either thinking through questions or giving up were relatively better options.

Incentivized website 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.[16] The web server tracked time spent on each activity (across all attempts) by recording a timestamp for each unique page view. Since only one math problem appears per page view within each activity, this resulted in a high-frequency log of work times for each child.[17] The website logged successful completions into a database, and included prominent visual indications to the user of completed tasks, piece-rate incentives, and current earnings. The final result of the website's meticulous tracking capability is that we could obtain data, on not just total website time T_i and task accomplishment A_i, but we also get 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 devices and internet connections at home. Roughly three quarters of pageload requests originated from computers (including laptops), about a fifth were from tablet devices, and about one in twenty pageloads were from smartphones. One potential concern is that limited internet access may have unduly influenced our results by inhibiting participation for some students. To address this concern, we collected survey data on internet connectivity at home. Although 7% of students reported not having a regular internet connection at home, these students were not statistically less likely to participate on the website than other

[15] Usernames and passwords 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. These turned out to be few, given the intuitive nature of the login credentials.

[16] The website also included an instructive component built from math textbook glossaries (generously furnished by the University of Chicago School Math Project, `ucsmp.uchicago.edu`) and practice materials by state boards of education. It contained glossary terms organized by math topic and a number of guided, interactive examples chosen to be representative of the paid materials on the site. This instructive component was clearly marked as non-incentivized to users, but it provided an option for students to invest in their income generation capability. However, less than 2% of overall page-view time was logged on the instructive portion of the website.

[17] 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 spurious work-times with student-sub-category-problem mean work times, using a procedure proposed by (Cotton et al., 2022, Online Appendix).

students, suggesting that they had other ways of accessing the internet.[18] We also found that having 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 a combination of 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 intuitive linear incentives with a constant piece-rate that would be easy for adolescents to understand: $\Pi_i(a_i) = (\pi_0^i + \pi_1^i a_i)\mathbb{1}(a_i \geq 2)$. Students were informed they must complete at least 2 website learning tasks (ensuring a within-student panel of data) in order to receive any payment. Each child was individually assigned at random to one of three incentive treatments, $(\pi_{01}^*, \pi_{11}^*) = (\$15, \$0.75)$, $(\pi_{02}^*, \pi_{12}^*) = (\$10, \$1.00)$, and $(\pi_{03}^*, \pi_{13}^*) = (\$5, \$1.25)$, thus ensuring treatment variation within schools, grades, and classrooms.[19] Our experimental design purposefully focused on piece-rate incentives—the dominant form of academic effort compensation—and achieved incentive salience by prominently advertising to each user on the home page his/her piece-rate, total accrued earnings (with real-time updates), and total remaining potential earnings. The question of adequate subject motivation depends on the ratio of payoffs to work quantity for each learning task. Breaking incentivized tasks into small units was useful for encouraging participation and for facilitating precise panel-data inference on θ_{pi}. Note that our three incentive schemes vary widely in proportional terms: the contract-2 and contract-3 marginal piece rates were, respectively, 33.3% and 66.7% larger than the contract-1 piece rate.

Our randomization algorithm first separated students into race-gender-school-grade bins. Within each bin, it balanced on pre-test scores by ordering students according to their score and randomly assigning consecutive blocks of 3 similar-score students to contract groups 1, 2, and 3. The algorithm then repeated this process thousands of times, and selected the candidate assignment that minimized overall correlations between treatment status and balance variables. A balance table (Table OS.4) is provided in the Online Appendix. Our final treatment assignment was independent of all balancing variables. On the Monday after the pre-test, each student participant received a personalized letter in a sealed envelope containing login credentials, instructions for accessing the website, a tech-support email address, and their individual piece-rate incentive contract.[20] They were also promised prompt delivery of payments within 2 weeks following the end of the experiment (which actually happened).

[18] 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.

[19] Base payments varied inversely with marginal wage only to mitigate possible concerns of fairness on the part of participants; otherwise, they play little role in the model or analysis. A pilot study indicated an expected average output of $\approx$20 completed tasks per student, at which point total payments across all three contracts are equal.

[20] A potential concern is that perhaps students shared their login credentials with others. While it's impossible for us to directly verify this, there are good reasons to believe it did not occur. First, nearly all students in a grade cohort received login credentials($> 95\%$), so they would not likely have been willing to do work for someone else if they could do the same work for themselves for profit. Second, roughly 1/2 of the sampled students declined any extracurricular 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 learning tasks logged by the system and incremental gains in math proficiency (using pre-test/post test scores, see Tables 8 and 6). This relationship persists even after controlling for a wealth of other student observables, and is strongly suggestive that completed website tasks logged by students reflect their own work, and not someone else's.

The design of our incentives had several advantages that encouraged ample effort from students on the whole, enabling structural inference on motivation and productivity for a wide variety of student types. First, we incentivized successful completion of learning tasks rather than time spent on these tasks. This is consistent with actual school environments where students are typically rewarded or punished (by schools and parents) based on whether they complete homework assignments. Similarly, we incentivized short-run, at-home math practice (analogous to a short homework assignment) rather than long-term outcomes such as year-end grades, making the decisions faced by students 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, List, and Sadoff (2016)) has shown that incentives are more effective when rewards follow actions as soon as possible. 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 will click "submit" on a given attempt only if she believes she may pass, but even in real time she cannot be certain of how much additional time it will take until the website reports that her last attempt finally cleared the bar for passing. Moreover, failed attempts can still motivate students to exert additional effort to achieve the intended result (Berger & Pope, 2011).

3.2.6. *Experiment Timeline.* The experiment proceeded as follows. (1) Teachers disseminated parental information sheets and assent forms two weeks prior to the launch of our study. (2) Students received their own assent form, and participated in a in-class pre-test and survey administered by their teachers. (3) Students were randomly assigned a compensation scheme and then provided with information about the experiment, website, login credentials, 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 scheme. (5) Teachers administered a post-test and survey in class. (6) Payments were mailed out within two weeks of the post-test and survey.

3.2.7. *Classroom Tests and Surveys.* In addition to website activity logs, we also collected test score and survey data on students. These data are not needed to identify individual motivation and productivity parameters, but they enable analyses in Sections 5 and 6 where we enhance interpretability of structural parameters by decomposing student (θ_p, θ_m) type estimates, and measure their relation to test scores and proficiency improvements. As part of our study, teachers in participating classrooms administered mathematics learning assessments (separate 30-45 minute exam based on common core sample problems) as pre- and post-tests for the experiment. Immediately following the pre- and post-tests, students also completed surveys, which collected information on a myriad of individual factors, including attitudes, extracurricular activities, regular study time, availability of parental homework support. We also gathered socioeconomic indicators from the American Community Survey for each of the 160 (rounded to preserve anonymity) 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.

3.3. **Descriptive Analyses of Website Activity.** Table 1 displays descriptive statistics of math website activity. Moving forward it will be useful to define "active students" as those who completed at least two website learning tasks, "marginal students" as those who completed one task but not a second, and "inactive students" as those who did not complete any. Active students were 44.7% of the sample, marginal students made up 5.6%, and inactive students were 49.7%. The raw data depict striking heterogeneity: 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 all have medians well below the means, and standard deviations near or above the means. Overall, we had 749 active students in our study, who collectively completed 16,740 learning tasks (i.e., between 84,000 and 100,000 math problems solved correctly) over the course of roughly 30,000 attempts and 2,000 child-hours during our 10-day sample period.

To place these numbers in context, 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).[21] One possible threat to our identification strategy would be if students responded to the extracurricular financial incentives by neglecting their schoolwork in proportion to their website activity. In multiple conversations with administrators and teachers involved in the study, they universally reported a firm impression that students did not reduce the amount of assigned homework they were actually turning in 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 this difference was statistically insignificant (p-value= 0.156).

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)*						

Aside from acting as a robustness check, this result helps us to contextualize the magnitude of observed website activity. Assuming mathematics accounted for 25%–50% of daily homework time implies the

[21] We asked students on the pre-survey: "How many hours do you usually spend on homework on a typical weekday (Monday through Thursday)?," and then we asked the same question applied to "...a typical weekend day (Friday-Sunday)?" Responses were multiple choice: "a. None; b. Less than 1 hour per day; c. Between 1 hour and 2 hours per day; d. Between 2 and 3 hours per day; e. More than 3 hours per day," and we coded a.– e. as 0, 1, 2, 3, and 4 hours, respectively. For average daily time spent, we computed (4/7)×(weekday avg. daily hmwk time)+(3/7)×(weekend avg. daily hmwk time). We repeated both questions on the post-survey, but there we asked students to think about the previous two weeks, specifically. We then averaged across pre- and post-survey responses.

average (median) website math time per day among active students would have represented a substantial increase of 37%–74% (24%–48%) in daily math activity by time.[22] Active students reported 22.1% more daily homework time than their marginal/inactive counterparts, 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. Within the full sample 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 of time spent on our extracurricular math website, with a Spearman rank correlation of -0.265 (p-value 2.6×10^{-26}). These empirical results suggest that our website observables are connected to unseen differences across students that produce disparate academic day-to-day choices, and disparate outcomes over time.

Figure 1. Website Choices and Performance

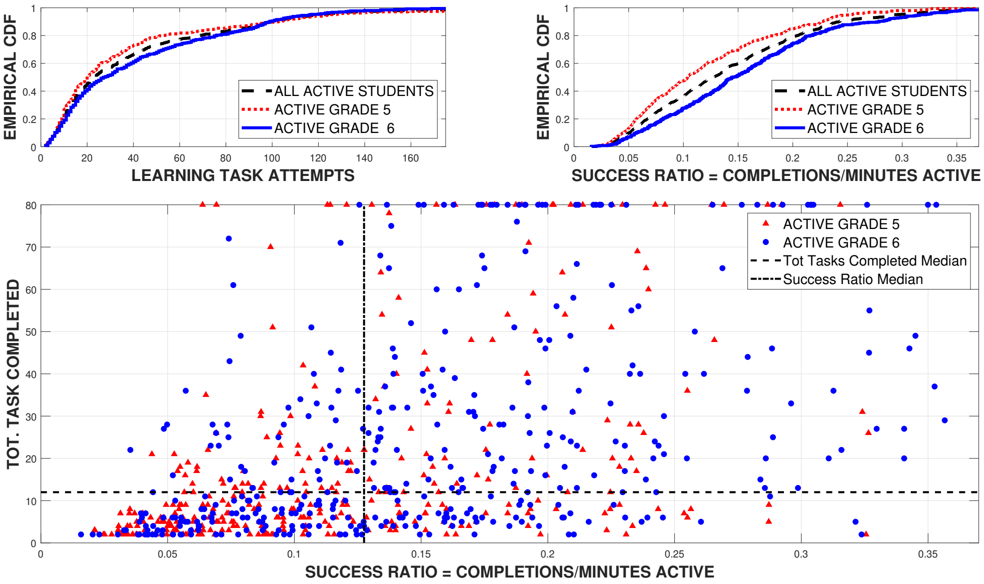


Figure 1 provides preliminary insights into unobserved student heterogeneity, based on field-experimental observables. The top two panels depict CDFs of student-level data on total learning task attempts, and what we refer to as the "success ratio" or task completions per unit of time, A_i/T_i. Both plots indicate vast heterogeneity, conditional on active status: the 90th percentile of total attempts is 17 times larger than the 10th percentile, and the 90th percentile child in terms of success ratio required only one fifth as much time to complete each task as the 10th percentile child. While the 5th and 6th grade CDFs are ordered as

[22] For an alternate benchmark, we discussed our findings on website math activity volume with a mathematics education consultant employed by a state board of education for a Midwestern U.S. state. Although volume of math problems assigned varies across classrooms, the consultant expressed the opinion that 72 extra math problems solved within a 10-day period (the median for active students) would be on par with an increase of between 50% and 100% in terms of regularly assigned homework volume for an average 5th or 6th grade student in the Midwestern United States.

one would expect, both groups have significant representation across a common support, providing further assurance that our choice to pool the two age cohorts within the experiment was reasonable.

The lower panel of Figure 1, a scatter-plot of success ratio A_i/T_i to task completion A_i, provides evidence on the relation between unobserved study productivity and motivation. The striking observation here is that the northwest quadrant (low success ratio but high task accomplishment) and the southeast quadrant (high success ratio but low task accomplishment) are both well populated. Many students who required $\geq$10 minutes per paid success completed well above the median number of learning tasks, while many other students requiring $\leq$5 minutes per paid success completed well below the median volume. The rank correlation between success ratio and task completion is (unsurprisingly) high at 0.551 (p-value 1.4×10^{-60}). On the other hand, the rank correlation between success ratio and website time choice is surprisingly low, at 0.067 (p-value 0.067). This reduced-form finding from our field experiment provides a new window into the "black box" of academic success, given that observational data on education typically provides only analogs of the vertical axis of the scatter plot (e.g., assignments turned in, exam scores, etc.).

A key prediction of our model is that sufficient strength on either trait, θ_p or θ_m, is enough to drive high observed performance. A very inefficient child (i.e., high θ_p) may still achieve at a high level with sufficiently high motivation (i.e., low θ_m), and vice versa. Figure 1 is strongly consistent with this idea. Conditional on learning task completion $A \geq 30$ (a standard deviation above the median) there is a striking degree of heterogeneity in completions per unit time. Two particular data-points in the scatter-plot, call them *child 1* $(0.057, 36)$ and *child 2* $(0.353, 37)$, vividly illustrate this point. While 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 math 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* still nearly matched *child 2*'s output, and therefore could be considered as *vastly more motivated.* Moreover, this begs the question of how much more *child 1* might have achieved under some intervention that narrowed the productivity gap between him/her and *child 2*, holding piece-rate incentives fixed.

4. Identification and Estimation of the Student Choice Model

The primary structural primitives of the study-leisure model are the idiosyncratic productivity (θ_{pi}) and motivation (θ_{mi}) parameters, and the common 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 structural identification and sketch out a two-stage GMM estimator to implement our identification strategy.

4.1. Stage-1 Estimation: Study-Time Productivity and Work-Time Shock Distributions.

Our approach here follows standard methods using the within-child panel structure of work-time data, $\{\tau_{a_i}\}_{a_i=1}^{A_i}$. Taking a log transformation of (1) gives a linear-in-parameters regression equation,

$$\log(\tau_{a_i}) = \log(\tau_0) + \log(\tau_1)\mathbb{1}(a_i\!=\!1) + \log(\theta_{pi}) - \varphi\log(a_i) + \log(u_{a_i}),\ \ a_i\!=\!1,\ldots,A_i,\ \ \left\{i\middle|A_i\!\geq\!1\right\}, \tag{5}$$

where productivity θ_{pi} enters as a student fixed-effect, and $(\tau_0, \tau_1, \varphi)$ enter as intercept and slope terms. We estimate regression parameters and fixed effects through a standard differencing approach.

For estimation of the heteroskedastic study-time shock distributions $F_u(u|\theta_p)$, we start by using regression estimates to compute fitted residuals $\hat{u}_{a_i}\!=\!\tau_{a_i}/(\hat{\tau}_0\hat{\tau}_1^{\mathbb{1}(a_i=1)}\hat{\theta}_{pi}a_i^{-\hat{\varphi}})$, $a_i = 1,\ldots,A_i$, $\{i|A_i \geq 1\}$. We allow for heteroskedastic shocks by partitioning the support of $\hat{\theta}_{pi}$ into 5 sub-intervals of equal length,

I_j^p, $j=1,\ldots,5$.[23] Then, we use this partition to split the sample of 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 $\hat{F}_u(u|I_j^p;\gamma_{uj})=\sum_{k=1}^{7}\gamma_{ujk}\mathcal{B}_{ujk}(u)$, $j=1,\ldots,5$.[24] Estimates are consistent with heteroskedastic work-time shocks: generally, students who take longer on average to solve math problems experience work-time shocks that are a mean-preserving spread relative to their quicker working counterparts.

All stage-1 model components can be separately pre-estimated under the following 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, the assumption 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 $\hat{F}_u(u|I_j^p;\hat{\gamma}_{uj})$, $j=1,\ldots,5$, can be treated as known (and fixed) during stage-2 estimation. This provides needed tractability by drastically reducing parameter-space dimension and computational burden, and by improving numerical stability for stage-2 estimates.

One challenge is that individual fixed-effect estimates have differing variances due to the unbalanced nature of the panel: A_i varies across active students, and higher values will lead to more precisely measured $\hat{\theta}_{pi}$.[25] Therefore, 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 some 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 are sufficient to nonparametrically identify a continuum of unobserved heterogeneity (θ_{mi} in our case) without *a priori* functional form restrictions on the (dis)utility function $c(t)$. Stage-2 identification relies on exogenous variation in observable incentivized actions, which our experimental design achieves via randomized incentive contracts (π_{0j},π_{1j}) across groups of adolescents, $j=1,2,3$, that are otherwise identical in their distributions of unobserved traits. Table 1 and Figure 2 show descriptive evidence of the exogenous data 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.

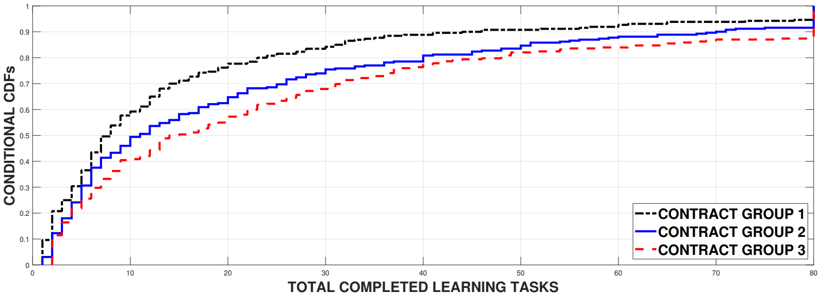
[23] Specifically, $I_j^p\equiv\left[\min(\hat{\theta}_{pi})+(j-1)h,\min(\hat{\theta}_{pi})+jh\right]$, $h=(\max(\hat{\theta}_{pi})-\min(\hat{\theta}_{pi}))/5$, $j=1,\ldots,5$. A finer partition of 10 sub-intervals of the support of θ_p made little difference in the following stages of estimation.

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

[25] Note that 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.

There are two basic tasks the Stage-2 structural estimator must accomplish with the experimental data: (i) pin down idiosyncratic labor-supply elasticities determined by θ_{mi}, and (ii) trace out the curvature of the labor-supply cost function $c(t)$. The conditional CDFs plotted in Figure 2 represent primary data moments relevant to these two tasks, and 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 1, 2, and 3. 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 depicted in Figure 2. For example, 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 (implemented using 100,000 bootstrap samples) in favor of the alternate hypothesis of first-order dominance with the following p-values: *Group 1* versus *Group 2*, p-value=0.002; *Group 1* versus *Group 3*, p-value$<10^{-5}$; *Group 2* versus *Group 3*, p-value=0.064.

Second, the figure depicts two unequal stochastic shifts. For task (ii), recall that the marginal wage difference between contracts 1 and 2 was the same as the difference between contracts 2 and 3: a \$0.25 raise per completed learning task. This fact, and strict cost curvature (i.e., $c''(t) > 0,\ \forall t$) produces the model prediction that quantile differences in work volume when shifting from contract 1 to contract 2 should generally be larger than those induced by the shift from contract 2 to contract 3. The data confirm this prediction: the 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 integrated 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. This is a 33% reduction in labor-supply response for the same level increase in the marginal piece-rate. This double difference helps pin down cost curvature.[26]

[26] As shown by Torgovitsky (2015) and D'Haultfoeuille and Février (2015, 2020), a single exogenous incentive shift may actually be sufficient to pin down the shape of $c(\cdot)$, as rich cost curvature information is encoded within the curvature of a single quantile difference $G_a^{-1}(r|\pi_{1j}) - G_a^{-1}(r|\pi_{1j'})$. However, the multiple cost shifts in our data-generating process improve statistical power and help to establish simpler intuition for the identification logic.

4.2.1. *Simulated GMM Estimator Overview.* Our GMM estimator is explicitly built upon functional representations of these counterfactual quantile comparisons. We start with a flexible cubic B-Spline specification of costs, $\hat{c}(t;\boldsymbol{\gamma}_c)=\sum_{k=1}^{k_C+3}\gamma_{ck}\mathcal{B}_{ck}(t)$, having knot vector $\boldsymbol{\kappa}_c=\{\kappa_{c1},\kappa_{c2},\ldots,\kappa_{c,K+1}\}$.[27]

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). To fix ideas, consider individual 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.[28] When solving for optimal stopping choices, we employ a new method recently proposed by Hamilton, Hickman, and Weidemann (2023) for tractable computation of non-stationary dynamic programming problems with history dependence.

Observed choices are also informative on cost curvature. First, A_i contributes to the empirical CDF of work volume under child 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 simulated values depend again on $F_u(u|\theta_{pi})$, θ_{pi}, τ_0, τ_1, and φ (all known and fixed), and on the shape of the cost function $\hat{c}(\cdot;\boldsymbol{\gamma}_c)$ through equations (1)–(4).

These 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, assuming child i was assigned to Contract Group 1, her observed choices (T_i, A_i) contribute to the empirical achievement CDF of her actual group, and they also contribute to the model-generated CDFs under counterfactual contract assignments 2 and 3. Of course, there is nothing special about a student being in Contract Group 1, and similar logic can be applied to all active students across all three contract groups. Cost function parameters $\boldsymbol{\gamma}_c$ are pinned down match child i's counterfactual projections to those of children at quantile rank r_i in contract groups 2 and 3 who, by random assignment, have similar characteristics (on average) to child i. More formally, the cost parameter estimator $\hat{\boldsymbol{\gamma}}_c$ is defined by minimizing the distance between the empirical CDFs $\hat{G}_a(\cdot|\pi_{1j})$, $j=1,2,3$, and their model-generated counterparts, $\tilde{G}_a(\cdot|\pi_{1j};\boldsymbol{\gamma}_c)$, $j=1,2,3$.

Bringing all of the above steps together, we obtain the following GMM objective function

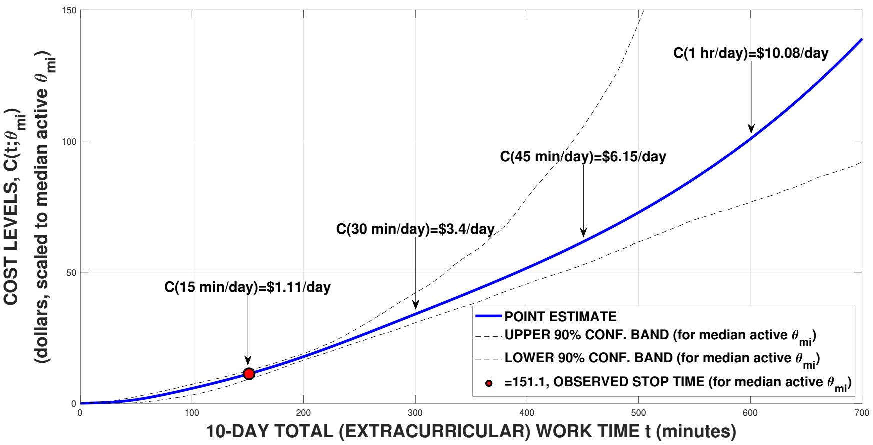
$$
\begin{aligned}
\hat{\boldsymbol{\gamma}}_c=\operatorname{argmin}\Bigg\{&\sum_{l=1}^{L}\sum_{j=1}^{3}\left(\hat{G}_a(a_l|\pi_{1j})-\tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right)^2\\
&+\omega_0\times\left(\hat{\overline{G}}{}_a^{90}(a_l|\pi_{1j})-\tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right)^2\times\mathbb{1}\left[\hat{\overline{G}}{}_a^{90}(a_l|\pi_{1j})<\tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right]\\
&+\omega_0\times\left(\underline{\hat{G}}{}_a^{90}(a_l|\pi_{1j})-\tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right)^2\times\mathbb{1}\left[\underline{\hat{G}}{}_a^{90}(a_l|\pi_{1j})>\tilde{G}_a(a_l|\pi_{1j};\boldsymbol{\gamma}_c)\right]\Bigg\}
\end{aligned}
\tag{6}
$$

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

[27] For the cost function we chose a knot vector with $K_c=7$ sub-intervals, or 8 knots spaced uniformly in quantile-rank space of T in order to evenly spread the influence of data over the various parameters $\boldsymbol{\gamma}_c$. After imposing the two boundary conditions in Assumption 2, this left 8 free parameters to allow the model-generated CDFs of A_i to fit their empirical analogs.

[28] 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.

Figure 3. Time Supply Cost & Marginal Cost Estimates



All CDF values are linearly interpolated on a grid $\{a_1, a_2, \ldots, a_L\} \subset [2, 80]$. The first line of the objective contains 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 $\underline{\hat{G}}_a^{90}(a_l|\pi_{1j})$ are the (interpolated) point-wise 90% confidence bounds of the empirical CDFs, and ω_0 is a penalty parameter. Guardrail conditions impose 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. This helps the solver to avoid becoming stuck at local optima $\boldsymbol{\gamma}_c$-space. The constraints enforce boundary value ($c(0) = 0$), boundary derivative ($c'(0) = 1$), monotonicity ($c'(t) > 0$), and convexity ($c''(t) > 0$), in that order.

4.2.2. *Correcting for Sample Selection.* There are two complications regarding mass points at the extremes of the sample that the simulated GMM estimator must also address. First, we have a small mass of students who achieve full output $A_i = 80$ on the website (see Figure 2). This issue is straightforward to deal with. Since these students' productivity types θ_p are known and precisely estimated, 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 challenge relates to sample selection issues in the upper tail of the (Θ_p, Θ_m) distribution, for the mass of students who did not complete enough tasks to get paid, $A_i \leq 1$. This is not a threat to identification of the structural model, which relies only on exogenous incentive variation across two groups that are the same on unobservable dimensions. Recall from Table 1 that the masses of active students, $M_j^{act} \equiv N_j^{act}/N_j$ in the three 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 "missing" students who would have entered active status ($A_i \geq 2$) under contract-3 incentives. For these two groups, we include marginal students (i.e., $A_i = 1$) within the simulated GMM routine described above, 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 $\tilde{G}_a(a|\pi_{1j})$. This ensures that the empirical CDFs $\hat{G}_a(A|\pi_{1j})$, $a \geq 2$, and their model-generated counterfactual analogs are based on underlying sets of students with comparable unobserved types.

4.3. **Structural Estimates.** Common Structural parameter estimates and bootstrapped confidence intervals are in Table 9 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 directly interpretable structural primitive is $\hat{\varphi} = 0.0788$, the experience-curve parameter. This estimated value is highly statistically significant (i.e., different from zero), but it implies only minor short-term productivity gains. Specifically, for any baseline of current work volume a_i, the mean per-unit completion time on the $(2a_i)^{\text{th}}$ task, $E\left[\tau_i(2a_i;\theta_{pi})\right]$ (i.e., after doubling work volume), drops by 5.32%.

4.3.1. *Cost Schedule.* Figure 3 plots the estimated cost function $C(T;\hat{\theta}_m,\widehat{\gamma}_c)$, scaled to the median value of θ_{mi} among active students. Costs and marginal costs are precisely estimated for relatively low values of time commitment, while the 90% bootstrapped confidence bands widen for higher values where time-choice data are sparse. Figure OS.3 in the 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 find that a high degree of curvature in the common cost function $c\left(t;\widehat{\gamma}_c\right)$ is required to rationalize the observed distributions of work activity. 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 per day would increase utility costs by an order of magnitude.

These numbers mask some subtle agency issues that exist within the educational context. Note that the labeled cost levels represent monetary transfers that would exactly offset utility costs associated with a certain commitment to marginal math activity beyond status-quo schoolwork. A principal who can force this median student into an extra hour of math study per day (if such a thing were needed to reach competency standards) could make the child whole again (i.e., zero surplus) with a daily transfer of \$10.08. However, this hypothetical assumes both access to the child's private information and a means of compelling him/her to the desired level of effort increase. Otherwise, the principal must simply offer incentives and allow the child to optimize, in which case he/she will choose an optimal stopping time that ensures strictly positive surplus. For example, the child depicted happened to be in contract group 3, and completed 28 learning tasks in 151.1 minutes, resulting in a surplus of around \$28. In that sense, the cost levels depicted are actually deceptively low: In order for the median student to rationally choose an additional hour of daily math study under private information and limited commitment, the principal would have to offer daily incentives far in excess of \$10.08.

4.3.2. *Motivation and Productivity Heterogeneity.* Figure 4 illustrates the degree of cost variation *across students.* The figure depicts cost schedules scaled to θ_m types at the 10th percentile (i.e., highly motivated), median, and 90th percentile (i.e., less motivated) of active students, where we have log-transformed costs to facilitate a graphical comparison. Here we see dramatic heterogeneity in willingness to supply time to math 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 90th-percentile θ_m type to study more will elicit very

Figure 4. Motivation Heterogeneity

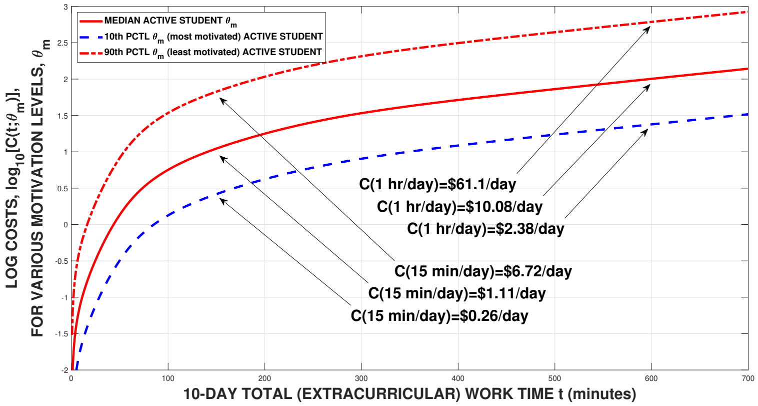
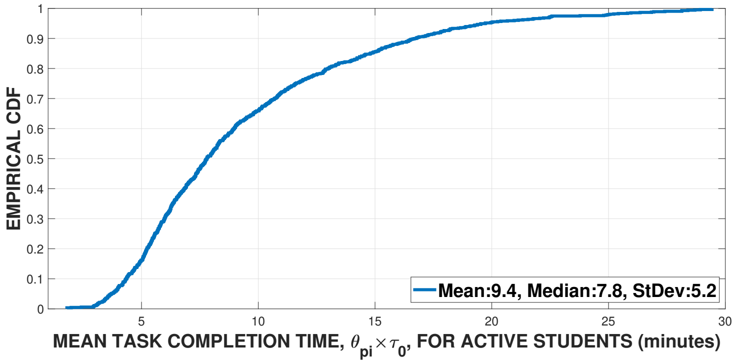


Figure 5. Productivity Heterogeneity



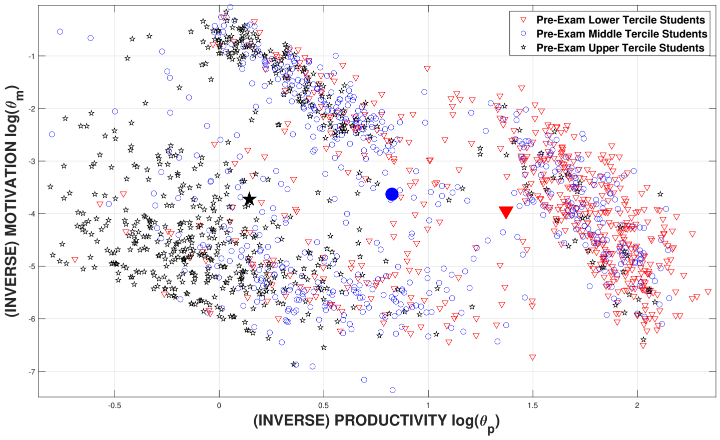
large (and very costly) responses by students who are much more motivated. On the other hand, providing lower uniform incentives that are only sufficient to entice the 90$^{\text{th}}$-percentile types will evoke little or no labor-supply response from the rest of the population. Of course, under piece-rate academic incentives, motivation, or willingness to supply a fixed quantity of time to study, is only one piece of the puzzle of student choice. Figure 5 depicts estimated productivity differences, where the 10-90 range entails more than a 4-fold increase in mean task completion times.

4.3.3. *Do low-performing students lack motivation?* Figure 6 jointly plots each student's productivity and motivation parameters, illustrating a wide range of type combinations. The Figure distinguishes students based on their pre-test score tercile. In the graph, the large star, circle, and triangle denote the average motivation and productivity values for each performance tercile. The structural analysis shows a wide variation in productivity-motivation combinations across students. Notably, we see a small negative correlation between $\log(\theta_p)$ and $\log(\theta_m)$, suggesting that students who take longer to successfully complete assignments may actually be more motivated than those who complete assignments more quickly, on average.

Similarly, when we compare the average productivity and motivation levels by pre-test performance tercile, we see very little difference in motivation across the three groups.[29] In fact, the lowest performing group tends to be slightly more motivated than their higher performing peers; although this difference is

[29] See Section 6 and Appendix A for more details on the pre- and post-tests.

Figure 6. Student Productivity and Motivation by Pre-Test Tercile



Notes: For active students ($A_i \geq 2$) plotted points represent Bayes shrunk forecasts of structurally estimated characteristics based on observed choices. For marginal/inactive students ($A_i < 2$) plotted points are estimated by integrating over the conditional mean of the non-identified region in $(\log(\theta_p), \log(\theta_m))$-space, given covariates combined with the parameters in Tables 2 and 3 (specification 4). The large bold star, circle, and triangle are mean $(\log(\theta_p), \log(\theta_m))$ values for students with pre-test scores in the upper tercile, middle tercile, and lower tercile, respectively.

not statistically significant. Both trends support the same conclusion: *lower performing students are not on average less motivated than higher performing students.* If anything, they may be slightly more motivated. This is true regardless of whether performance is judged on the completion of assignments or performance on proficiency assessments.

Where we see the biggest differences between higher and lower performing students is in terms of their productivity parameters. Students in the higher performing groups require substantially less time to complete a given amount of homework than their classmates, allowing them to work through more assignments in a given amount of time. In the following subsections, we take a detailed look at the drivers of productivity and motivation, and explore how these factors interact with effort as drivers of learning.

5. DECOMPOSITION OF MOTIVATION AND PRODUCTIVITY

The previous section estimated individual-level motivation (θ_{mi}) and study-time productivity (θ_{pi}) traits. In this section we explore sources of productivity and motivation heterogeneity.

5.1. **Decomposition of Student Heterogeneity.** Idiosyncratic differences in student motivation traits θ_m may be driven by either opportunity costs of foregone leisure time, the quality and variety of outside options, or by direct psychic costs of working through math problems. Heterogeneity in productivity θ_p may reflect either foundational cognitive or non-cognitive differences, initial proficiency level with the current-grade concepts, or differences in a child's study process, academic support network, school quality, or other environmental factors. Since both traits are a mixture of innate and environmental components, for each student i we allow them to depend on changing circumstances as follows:

$$\log(\theta_{pi}) = \boldsymbol{X}_{pi}\boldsymbol{\beta}_p + \eta_{pi}, \quad \text{and} \quad \log(\theta_{mi}) = \boldsymbol{X}_{mi}\boldsymbol{\beta}_m + \eta_{mi}, \tag{7}$$

where $\boldsymbol{X}_{pi} = \left[1, x_{p1i}, \ldots, x_{pk_{p}i}\right]$ and $\boldsymbol{X}_{mi} = [1, x_{m1i}, \ldots, x_{mk_{m}i}]$ are vectors of student-level covariates, and the η_{pi} and η_{mi} terms represent the truly idiosyncratic portions of student i's unobserved traits $(\theta_{pi}, \theta_{mi})$. The covariate vector, $\boldsymbol{X}_{pi}$, for the study-time productivity equation contains an intercept term and the following variables: indicators for *gender*, *race*, *grade level*, and *school district*; the *# of adult academic helpers* in a child's social network; the *# of peer academic helpers*; and two socioeconomic proxies specific to the child's neighborhood of residence: *mean household income* (a proxy for affluence), *fraction of minors with no private health insurance*, and a dummy variable for *no home internet connection* (both proxies for deprivation of non-school developmental resources). The covariate vector $\boldsymbol{X}_{mi}$ for motivation contains these same variables and adds an additional set of variables pertaining specifically to attitudes, preferences, and outside options for time use, including indicators for whether *math is a favorite* academic subject or *math is a least favorite* subject; *extrinsic motivation score*; *intrinsic motivation score*; indicators for enrollment in organized *sports*, organized *music* activities, other organized *clubs*; *fraction of peer social time under adult supervision*; *# of video gaming systems* at a child's home; *parental permission for video gaming on weekdays*; *weekday time spent on recreational internet use*; *mean daily recreational screen time*; and *mean daily regular schoolwork time.* The idea in adding these additional factors to equation (7) is that θ_{mi} represents a child's level of motivation for shifting an hour of her time away from the best outside option (e.g., gaming, internet surfing, playing with friends) and toward math activity, which may be influenced by her attitude toward math or her responsiveness to different forms of incentives, holding her study-time productivity θ_{pi} fixed. These variables are all summarized in Table 8 in Appendix A.3.

The challenge here is a basic sample truncation problem: while $(\boldsymbol{X}_{pi}, \boldsymbol{X}_{mi})$ is known for all $i = 1, \ldots, N$, the outcome variables $(\log(\theta_{pi}), \log(\theta_{mi}))$ are known only for students who chose $A_i \geq 2$. So solve this issue, we adopt a parametric assumption on the underlying idiosyncratic types:

Assumption 5. Residual productivity and motivation heterogeneity (η_p, η_m) follow a bivariate normal distribution, $(\eta_p, \eta_m) \sim BVN(\mathbf{0}, \boldsymbol{\Sigma})$, where the covariance structure may vary by race and gender: $\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma^2_{pi}\sigma_{pli} \\ \sigma_{pli}\sigma^2_{mi} \end{bmatrix}$, and $\sigma_{ji} = \sigma_{j0} + \sigma_{j1} fem_i + \sigma_{j2} black_i + \sigma_{j3} hispanic_i, \quad j = p, m, pm.$

By adopting Assumption 5, we can implement a 2-dimensional Maximum Likelihood Tobit strategy, using the known, contract-specific selection thresholds $\underline{\Theta}_m(\theta_p; \pi^*_{0k}, \pi^*_{1k}, \widehat{\gamma}_c)$, $k = 1, 2, 3$, which can be estimated as the northeast boundary of the convex hull of identified types uncovered in the previous stages of estimation. Our Tobit estimator is thus defined by optimizing the following log-likelihood function:

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

where the ω_{di}'s are inverse-variance weights: $\omega_{di} = \frac{1}{Var(\hat{\theta}_{pi})}$ whenever $A_i \geq 2$, and $\omega_{di} = \min\{\omega_{dj} | A_j \geq 2\}$ whenever $A_i < 2$. For tractability, we compute the probability in the Tobit term above by simulation. We use the bootstrap method in combination with bootstrapped structural estimates in order to adjust standard errors for sampling variability from previous stages of estimation.

5.2. **Empirical Results: Student Type Decomposition.** Figures 7 and 8 illustrate the distributional differences in θ_p, θ_m, pre-test score, and number of learning tasks completed by race and gender. Tables 2

Figure 7. Distributions of Characteristics by Race

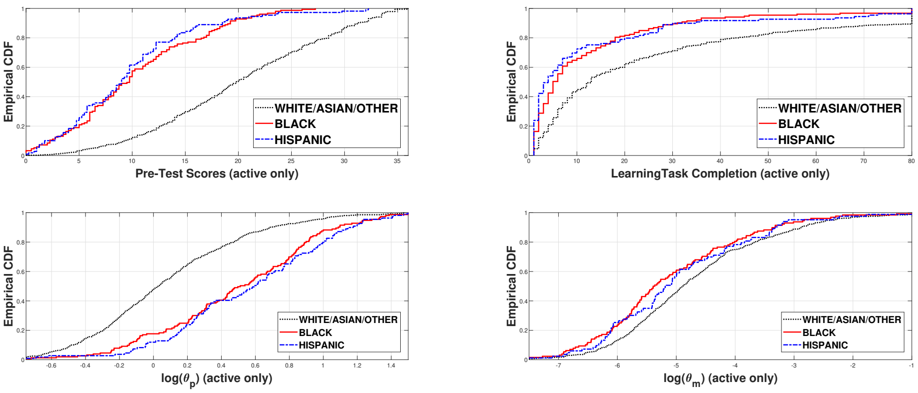
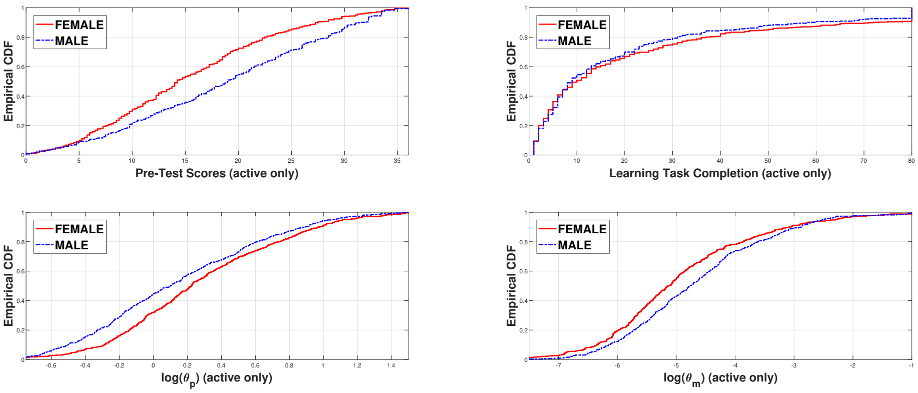


Figure 8. Distributions of Characteristics by Gender



and 3 report results from Tobit regressions exploring relationships between observable student characteristics and the motivation and productivity parameters from the previous section. Although our model is primarily one of short-term choices, our pooling of 5thand 6thgraders 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 5th-graders and 6th-graders are indistinguishable, on average, in terms of their motivation for math study, but 5th-graders are less productive by 30% of a standard deviation after controlling for the full set of student covariates. This last result is presumably the impact of an additional year of instructional inputs and learning activity on 5th-/6th-grade math materials, which are very similar in content (see Online Appendix B).

Another significant lesson from Table 3 concerns external validity of our motivation index θ_p derived from the revealed-preference principal using our field-experimental incentives for extracurricular math activity. Recall from Section 3.1.1 that a potential limitation of our method may be that it merely measures willingness to allocate *extra time* to math learning, above and beyond regular schoolwork. A lingering question therefore is whether estimated motivation heterogeneity represents deeper motivational differences across students, or whether it is merely reflecting differences in baseline coursework load and differing levels of associated burnout. Our rich outside time-use data from student surveys allow us to directly address this concern, which would imply that the coefficient on *Reg. Study Time* in Table 3 should be positive. In other

Table 2. TOBIT REGRESSION RESULTS: STUDY-TIME PRODUCTIVITY

SPECIFICATION:	**(1)**		**(2)**		**(3)**		**(4)**	
DEP. VAR.: $\log(\theta_p)$	Estimate	StdDev Effect	Estimate	StdDev Effect	Estimate	StdDev Effect	Estimate	StdDev Effect
Female ($\widehat{\beta}_{p1}$)	-0.093	-0.066	-0.124	-0.104	0.208***	**0.201**	0.192***	**0.205**
(std. err.)	*(0.116)*		*(0.100)*		*(0.039)*		*(0.038)*	
Black ($\widehat{\beta}_{p2}$)	0.990***	**0.704**	0.895***	**0.752**	0.964***	**0.931**	0.942***	**1.008**
(std. err.)	*(0.146)*		*(0.144)*		*(0.079)*		*(0.113)*	
Hispanic ($\widehat{\beta}_{p3}$)	1.131***	**0.804**	0.776***	**0.652**	0.955***	**0.922**	0.723***	**0.774**
(std. err.)	*(0.177)*		*(0.205)*		*(0.171)*		*(0.127)*	
Grade 5 ($\widehat{\beta}_{p4}$)	0.362***	**0.258**	0.250***	**0.210**	0.260***	**0.251**	0.280***	**0.300**
(std. err.)	*(0.058)*		*(0.051)*		*(0039)*		*(0.044)*	
District 2 ($\widehat{\beta}_{p5}$)	—	—	0.203**	**0.171**	0.167***	**0.162**	0.149**	**0.159**
(std. err.)			*(0.094)*		*(0.080)*		*(0.084)*	
District 3 ($\widehat{\beta}_{p6}$)	—	—	1.000***	**0.840**	0.713***	**0.689**	0.632***	**0.677**
(std. err.)			*(0.197)*		*(0.179)*		*(0.166)*	
No Home Internet ($\widehat{\beta}_{p7}$)	—	—	—	—	—	—	0.399***	**0.427**
(std. err.)							*(0.143)*	
Constant ($\widehat{\beta}_{p0}$)	0.296		0.110		-0.300		-0.322***	
(std. err.)	*(0.119)*		*(0.121)*		*(0.004)*		*(0.003)*	
Nbhd. SES	YES		YES*		YES		YES	
Fam. Support	no		no		YES*		YES**	
N	1,676		1,676		1,676		1,676	
Pseudo-R^2	0.232		0.378		0.474		0.473	
log-Likelihood	-3,501.4		-3,498.9		-3,411.0		-3,386.2	

Notes: ***Higher* dependent variable values** $\log(\theta_p)$ **imply *lower* study-time productivity. Nbhd. SES** controls serve as proxies for socioeconomic resources, including (standardized) log of mean income and (standardized) fraction of minors with no health insurance within the US Census blockgroup where the child resides. **Fam. Support** controls include (self-reported) counts of how many adults (e.g., parent, tutor, etc.), and how many peers (e.g., friend, sibling, cousin, etc.) regularly help the student with his/her math homework. **StDev Effect** represents the change in standard deviation units of $\log(\theta_p)$ from 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 the highest statistical significance level for a single variable within that group. In all Nbhd. SES and Fam. Support controls play a minor role in explaining math study-time productivity. Due to joint Tobit Estimation, **Pseudo-R^2** for $\log(\theta_p)$ need not increase monotonically with model richness.

words, students who are estimated to be less motivated within our study (i.e., high $\log(\theta_m)$) may tend to be those who are more motivated and committed to regular coursework duties and thus log more regular study time (i.e., positive $\hat{\beta}_{m11}$). 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. However, this potential concern is not supported by the data: the coefficient estimate $\hat{\beta}_{m11}$ is actually negative and statistically different from zero at the 95% level. This means that students who are more committed to regular coursework, are also more committed to extracurricular math activity as well, after controlling for the full set of 24 student-level covariates. Although the effect is small, it increases confidence in our field-experimental and structural methodology as tapping into underlying factors that drive academic choices on a day-to-day basis in real students' lives.

5.2.1. *Racial differences.* From Figure 7 and specification (1) of both regression tables, we observe that Black and Hispanic students are on average moderately more motivated and substantially less productive than White/Asian students. This means that while Black/Hispanic students are on average more willing to spend an hour studying, it also takes them more hours to complete a given set of learning tasks. The productivity disadvantage dominates their motivational advantage, and on net Blacks and Hispanics tend to complete fewer website learning tasks than White/Asian students in our sample.

Productivity differences are robust to controlling for gender, school district, and other factors that reflect resources like peer/adult support, socioeconomic proxies, and home internet access. The racial gap is

Table 3. TOBIT REGRESSION RESULTS: MOTIVATION

SPECIFICATION:	**(1)**		**(2)**		**(3)**		**(4)**	
DEP. VAR.: $\log(\theta_m)$	Estimate	StDev Effect	Estimate	StDev Effect	Estimate	StDev Effect	Estimate	StDev Effect
Female ($\widehat{\beta}_{m1}$)	-0.368	-0.244	0.486	0.315	-0.695***	**-0.337**	-0.721***	**-0.344**
(std. err.)	*(0.526)*		*(0.344)*		*(0.115)*		*(0.101)*	
Black ($\widehat{\beta}_{m2}$)	-1.260***	**-0.834**	-0.583	-0.378	-1.449***	**-0.702**	-1.425***	**-0.679**
(std. err.)	*(0.506)*		*(0.403)*		*(0.364)*		*(0.374)*	
Hispanic ($\widehat{\beta}_{m3}$)	-0.917*	**-0.606**	-0.706	-0.458	-1.078*	-0.522	-0.365	-0.174
(std. err.)	*(0.598)*		*(0.558)*		*(0.584)*		*(0.444)*	
Grade 5 ($\widehat{\beta}_{m4}$)	-0.300**	**-0.198**	-0.136	-0.088	-0.012	-0.006	-0.093	-0.044
(std. err.)	*(0.151)*		*(0.158)*		*(0.124)*		*(0.117)*	
District 2 ($\widehat{\beta}_{m5}$)	—	—	0.001	0.001	0.553**	**0.268**	0.543***	**0.259**
(std. err.)			*(0.287)*		*(0.264)*		*(0.234)*	
District 3 ($\widehat{\beta}_{m6}$)	—	—	0.190	0.124	0.675	0.327	0.801**	**0.382**
(std. err.)			*(0.779)*		*(0.541)*		*(0.409)*	
Math Favorite ($\widehat{\beta}_{m7}$)	—	—	—	—	-0.300***	**-0.145**	-0.269**	**-0.128**
(std. err.)					*(0.150)*		*(0.148)*	
Math Least Fav. ($\widehat{\beta}_{m8}$)	—	—	—	—	0.267	0.130	0.335*	**0.159**
(std. err.)					*(0.200)*		*(0.200)*	
Intrinsic Mindset Score ($\widehat{\beta}_{m9}$)	—	—	—	—	-0.527***	**-0.229**	-0.570***	**-0.244**
(std. err.)					*(0.128)*		*(0.107)*	
Extrinsic Mindset Score ($\widehat{\beta}_{m10}$)	—	—	—	—	-0.520***	**-0.211**	-0.638***	**-0.255**
(std. err.)					*(0.117)*		*(0.110)*	
Reg. Study Time ($\widehat{\beta}_{m11}$)	—	—	—	—	—	—	-0.179**	**-0.058**
(std. err.)							*(0.070)*	
Rec. Screen Time ($\widehat{\beta}_{m12}$)	—	—	—	—	—	—	0.146*	**0.085**
(std. err.)							*(0.072)*	
Constant ($\widehat{\beta}_{m0}$)	-4.690***	—	-5.216***	—	-3.360***	—	-3.404***	—
(std. err.)	*(0.321)*		*(0.197)*		*(0.005)*		*(0.004)*	
Nbhd. SES	YES**		YES		YES***		YES***	
Fam. Support	no		no		YES		YES	
Extracurriculars	no		no		YES		YES*	
Gaming & Surfing	no		no		YES		YES	
No Home Internet	no		no		no		YES	
N	1,676		1,676		1,676		1,676	
Pseudo-R^2	0.201		0.226		0.372		0.360	
log-Likelihood	-3,501.4		-3,498.9		-3,411.0		-3,386.2	

Notes: ***Higher*** **dependent variable values** $\log(\theta_m)$ **imply *lower* motivation. Nbhd. SES** controls serve as proxies for socioeconomic resources, including (standardized) log of mean income and (standardized) fraction of minors with no health insurance within the US Census blockgroup where the child resides. **Fam. Support** controls include (self-reported) counts of how many adults (e.g., parent, tutor, etc.), and how many peers (e.g., friend, sibling, cousin, etc.) regularly help the student with his/her math homework. **Extracurriculars** controls include dummy variables for enrollment in sports, music, and clubs, as well as fraction of social time in structured, adult-supervised activities. **Gaming & Surfing** controls include # of video gaming systems at a student's home, and parental permission for playing video games or recreational internet use on weekdays. **StDev Effect** represents the change in standard deviation units of $\log(\theta_m)$ from switching a binary regressor value from 0 to 1, or from increasing the value of a continuous regressor 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 the highest statistical significance level for a single variable within that group. Due to joint Tobit Estimation **Pseudo-R^2** for $\log(\theta_p)$ need not increase monotonically with model richness.

substantial, with $\log(\theta_p)$ being on average roughly 1 SD higher for Black students and 0.77 SD higher for Hispanic students, relative to their White/Asian classmates, after controlling for socioeconomic proxies and other observable factors in specification (4). To put this in context, in Specification (4) we find that the SD productivity advantage of White/Asian students relative to their minority peers is between 2.6 and 3.4 times the productivity advantage associated with an additional year of schooling. The statistical significance of the Hispanic motivation advantage is less robust to controlling for observable factors, particularly school district, in Table 3. However, despite losing significance in some specifications, the coefficients (all negative) consistently show that minority groups tend to be more motivated than their White/Asian peers. After controlling for the full set of 24 student covariates in Specification (4), including attitudes towards math, typical time use, family background, etc., we find statistically significant evidence that Black students on average are more motivated than non-black peers by roughly two-thirds of a standard deviation.

5.2.2. *Gender differences.* Figure 8 illustrates a moderate female advantage in terms of motivation and a moderate female disadvantage in terms of study-time productivity. From the tables, however, we see that neither of these gender gaps are statistically significant before controlling for a richer set of observable student characteristics, including whether the students have family members whom they rely on for homework support and other factors. In the later productivity regressions, we find that $\log(\theta_p)$ is approximately 0.21 SD higher for females than males, suggesting that females on average take longer to complete a given set of homework assignments. Although the gender gap is substantially smaller than the racial gap, it is not negligible in the later specifications. In specification (4) of Table 2, the male study-time productivity advantage is approximately 2/3 the productivity advantage associated with one year of schooling. However, in specification (4) of Table 3, we find strong evidence that female adolescent students are also more motivated, with $\log(\theta_m)$ being 1/3 SD lower, on average, for females than for males. On net, the latter effect dominates and in Figure 8 we see females on average completing more website learning tasks.

5.2.3. *Differences across school districts.* Now we turn to the role of school quality in shaping adolescent productivity and motivation. In Table 2, after controlling for observable student characteristics, one's school enrollment predicts significant reductions in time required for a student to complete learning tasks. From the descriptive evidence in Table OS.3, one might have suspected that District 1's inputs—higher funding per student, larger fraction of budget devoted to instruction, and better paid faculty/administrators—and its performance record—higher fraction of students meeting/exceeding state expectations—are more advantageous to the student than District 2's, which are in turn more advantageous than District 3's. Although school-district covariates were not included in the present analysis, this pattern plays out in the value-added estimates from the Tobit model: switching from District 1 to District 2 or District 3 induces a reduction in a child's study-time productivity by 0.16 SD and 0.68 SD, respectively. The latter result is more than twice the gap between grade-5 and grade 6-students, holding school district and all other student observables fixed. Coefficient estimate magnitudes are stable across the last three specifications, suggesting robustness of this result to inclusion of a rich set of other childhood contextual factors.

Similar patterns emerge for motivation levels θ_m as well. In specification (4) of Table 3, switching from District 1 to District 2 or District 3 induces a statistically significant drop in motivation level by 0.26 SD and 0.38 SD, respectively. Once again we see robustness of these estimated magnitudes to inclusion of a rich set of other childhood contextual factors across specifications (3) and (4), including preferences, attitudes, consumption proxies, and outside options for time use, among others. Our Tobit results speak to a classic question of whether better outcomes at higher-performing schools are due primarily to treatment by more advantageous school inputs, or whether they are due to selection of more academically adept students onto their rolls. We indeed find strong evidence for both explanations of gaps in academic outcomes: higher-performing schools do benefit from significant advantageous selection on both θ_p and θ_m (see Figure OS.4, Online Appendix B), but at the same time they also causally contribute to the productivity and motivation differentials in a substantial way. In the following section, we further investigate whether/how schools produce value-added in the learning process by shaping production technology as well.

It is worth mentioning that our identification strategy for causal value-added differs significantly from existing methods, due to the unique set of observable student variables and exogenous variation that our field experimental design facilitated. A typical study on school value-added would use observational data with a large sample of schools, and outcomes (e.g., exam scores) often aggregated to the classroom or school level. A typical study would then appeal to some sort of plausibly exogenous variation to tease apart selection on unobserved student traits from causal school value-added (e.g., Dale and Krueger (2002) and Mountjoy and Hickman (2020) apply such techniques to colleges). In our case, we have a small set of school districts, but incredibly rich, student-level data, including behavioral responses to exogenously varying incentives that facilitate identification of unobserved student characteristics, independently of school characteristics.

This solves the canonical problem in school value-added of selection on student unobserved traits by allowing the researcher to directly quantify and explicitly control for these unobserved traits. This is why our study is able to derive plausibly causal school value-added estimates from a small sample of schools. Given structural estimates based on incentive variation which was independent of schools and teachers, causal interpretation of the value-added parameters $(\hat{\beta}_{p5}, \hat{\beta}_{p6})$ and $(\hat{\beta}_{m5}, \hat{\beta}_{m6})$ reduces to a simple question of omitted variable bias, a concern which is mitigated by our rich data and stability of estimated parameter magnitudes across multiple specifications in Tables 2 and 3.

5.2.4. *Other considerations.* Student self-reports of math being a favorite subject is predictive of a significant increase in willingness to spend time on math by 0.13 SD, while listing math as one's least favorite subject induces a significant reduction in willingness to allocate time to math by a similar amount, 0.16 SD. The presence of psychic costs of academic effort has long been theorized within the related literature. These novel results contribute by directly quantifying aspects of psychic costs: we find that preferences for or aversion to the subject material indeed play a non-trivial role in utility costs incurred by math activity, holding all else equal. We also find that being either more intrinsically minded or more extrinsically minded are *both* strong indicators of responsiveness to our extrinsic financial incentives for students to divert extra leisure time toward math activity.[30] 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., 2019), rather than a conflicting role as previously thought (e.g., Gneezy & Rustichini, 2000; Bénabou & Tirole, 2003; Leuven, Oosterbeek, & van der Klaauw, 2010). The intrinsic mindset impact, 0.24 SD more motivation, also speaks to the role of psychic costs in determining effort: a highly intrinsic mindset is more than enough to overcome the drop in motivation from having math as one's least favorite academic subject.

We also assess the relationship between socioeconomics and the current values of θ_p and θ_m. We have two measures of the socioeconomic well-being of a student's census block group, including the log of mean neighborhood income and the share of minors without private health insurance. The first is a measure of affluence, while the second is a measure of developmental resource deprivation. While neither plays a meaningful role in determining productivity θ_p, we find suggestive evidence of a significant role for

[30] Intrinsic/Extrinsic mindset scores were derived as follows. We included two questions each on the pre-survey and post-survey asking students about their 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 then counted the number of corresponding responses across the four questions and standardize the score by subtracting means and dividing by standard deviations. Given the presence of the fifth option, it is possible for a student to coded as exhibiting significant levels of extrinsic mindset, intrinsic mindset, both, or neither.

motivation toward academic pursuits.[31] Although not explicitly reported in the tables, the dominant factor among neighborhood SES controls is the fraction of minors with no health insurance coverage. When this factor rises significantly above its mean by half of a standard deviation or more, the model predicts that neighborhood SES proxies on net contribute to a non-trivial degradation of willingness to spend time on math activity. This pattern is illustrated in Figure 16 (Appendix A). This finding points to developmental resource deprivation as a significant driver of a child being *less* motivated for schoolwork.

6. EXPLORING THE DETERMINANTS OF PROFICIENCY AND PROGRESS

The current section explores how structurally estimated student traits and other environmental factors contribute to student exam performance and measured proficiency gains. The framework we study here is illustrative of the inferential power to be had from being able to directly quantify underlying motivation and productivity. Similar logic applies here as in Section 5.2.3 for our causal school value-added estimates. We have a small set of school districts in our sample, but we base inference on incredibly rich, student-level data, including exogenous incentive shifts. This solves the canonical problem in school value-added of selection on student unobserved traits: given structural estimates $(\hat{\theta}_{pi}, \hat{\theta}_{mi})$ based on incentive variation which was independent of schools and teachers, causal interpretation of school value-added parameters reduces to a simple question of adequately controlling for relevant sources of omitted variable bias.

We first explore the determinants of initial mathematics proficiency (Table 5). The literature typically uses test scores to measure proficiency, and in this section we do the same, relying on student scores on the classroom pre-test, taken during the week before the website became available. Let S_i denote one's initial math proficiency measured by pre-test score. Second, this section explores the determinants of proficiency improvements. We measure progress as changes in classroom test scores between the pre-test before the experimental period and the post-test afterward, denoted ΔS_i (Table 6). In these models, we allow for student characteristics to not only determine intermediate inputs (T_i, A_i), but also to influence the rate at which child i converts a fixed volume of work into permanent proficiency gains as measured by the exam. Third, conduct model simulations and counterfactual analyses, exploring how differences in school-quality contribute to racial performance gaps and responsiveness to general academic incentives.

Table 4 shows descriptive statistics on average pre-test scores and proficiency gains by sub-group and the top left graphs in Figures 7 and 8 illustrate the pre-test score distributions by race and gender. The data highlight a substantial racial gap in test performance, which is generally consistent with evidence of substantial demographic mathematics gaps from other studies (e.g. Clotfelter, Ladd, & Vigdor, 2009; Hanushek & Rivkin, 2006, 2009; NAEP, 2019). White/Asian students performed substantially higher on the standardized mathematics pre-assessment than their Black and Hispanic peers, with the average White/Asian student correctly answering roughly 10 additional questions (1.13 SD), relative to the average minority student. The gender gap is relatively small compared to racial gaps in scores, with the average male correctly answering 1.4 more exam questions than the average female, corresponding to 0.16 SD higher performance.

The table also highlights how extracurricular math activity on our website during the sample period contributed significantly to measured proficiency gains. Active students saw increases of 2.67 exam questions

[31] A note of caution regarding interpretation: our socioeconomic controls are measured at the neighborhood (i.e., Census block group) level rather than at the household level, so this result may not represent the causal impact of health insurance *per se*, but should be regarded as a stand-in for endowment of non-school developmental resources.

TABLE 4. 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_1	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 (Post-Pre)	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: Unless otherwise stated, standard font numbers in the table represent sample means, while italicized numbers in parentheses represent sample standard deviations. **The null hypothesis that incentivized website activity did not result in learning gains, or $H_0 : E[\Delta Score|Active] = E[\Delta Score|Marg./Inactive]$, is firmly rejected by a two-sample t-test with a p-value of 1.2×10^{-17}.** Fifth-graders make up 47.3% of the total sample, with 6th graders comprising the other 52.7%. Sub-sample proportions are close to that ratio for all gender and race groups.

answered correctly, on average, while marginal/inactive students improved their scores by only 0.51 correct answers. Both changes are statistically significant at the 1% level, and the latter provides a useful baseline for the default learning that happens through regular coursework over a 2-week period.

6.1. **Determinants of mathematics proficiency.** We model initial math proficiency as the outcome of a Cobb-Douglass production process with θ_{pi} and θ_{mi} as its principal inputs, and where production shares and total factor productivity (TFP) terms are allowed to vary by individual i:[32]

$$S_i = TFP_i \times \theta_{pi}^{\alpha_{pi}} \times \theta_{mi}^{\alpha_{mi}} \times \epsilon_i, \quad TFP_i > 0, \;\; \alpha_{pi} < 0, \;\; \alpha_{mi} < 0. \tag{9}$$

TFP and the production shares $(\alpha_{pi}, \alpha_{mi})$ are not random coefficients, but are functions of covariates

$$\log(TFP_i) = \boldsymbol{W}_i\boldsymbol{\alpha}_0, \quad \alpha_{pi} = \boldsymbol{W}_i\boldsymbol{\alpha}_p, \quad \text{and} \quad \alpha_{mi} = \boldsymbol{W}_i\boldsymbol{\alpha}_m, \tag{10}$$

with $\boldsymbol{W}_i = [1, w_{1i}, \ldots, w_{ki}]$, including a constant and various student-level contextual factors. The error term ϵ_i is an idiosyncratic shock that accounts for cumulative impacts of transitory perturbations to HC production and noise in the exam instrument used to measure math proficiency.

One's score on the classroom pre-test, S_i, provides a baseline measure of skill stock, while productivity governs the rate at which intermediate learning tasks are traversed during one's efforts to augment skill stock. While the two are certainly related concepts, they are not the same thing. Many of the factors that influence S_i over the short-run also influence evolution of θ_p over the long-run, with both measures influenced by general aptitude, foundational skills, knowledge of relevant concepts, practice, and different aspects of attention or anxiety, for example. However, θ_p reflects the the amount of time one requires to correctly solve math problems outside of the classroom in an un-structured homework setting, given real-time feedback on incorrect answers, as well as access to textbooks, supporting materials, examples, and assistance from friends or family members. On the other hand, proficiency stock S_i is a measure of a child's

[32] When interpreting empirical results, 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.

performance taken in a controlled, timed, classroom exam environment, without any real-time feedback or access to external aids.

6.1.1. *Estimating the model.* Substituting equation (10) into equation (9), the initial proficiency model is equivalent to a regression of $\log(S_i)$ on θ_{pi}, θ_{mi}, $\boldsymbol{W}_i$, and a complete set of pair-wise interactions between $(\theta_{pi}, \theta_{mi})$ and $\boldsymbol{W}_i$:

$$\log(S_i) = \boldsymbol{W}_i\boldsymbol{\alpha}_0 + \boldsymbol{W}_i\boldsymbol{\alpha}_p \log(\theta_{pi}) + \boldsymbol{W}_i\boldsymbol{\alpha}_m \log(\theta_{mi}) + \log(\epsilon_i). \tag{11}$$

The covariate vector, $\boldsymbol{W}_i$, contains an intercept term and the following variables: indicators for *gender*, *race*, *grade level*, and *school district*; neighborhood-level socioeconomic indicators *mean neighborhood income* (a proxy for affluence) and *fraction of neighborhood minors with no private health insurance* (a proxy for deprivation of non-school developmental resources); and *total # of academic helpers* in a child's social network. Note that each of these factors is allowed to have a direct impact (through the intercept terms $\boldsymbol{W}_i\boldsymbol{\alpha}_0$), and also to have an indirect impact (through the slope terms $\boldsymbol{W}_i\boldsymbol{\alpha}_p$ and $\boldsymbol{W}_i\boldsymbol{\alpha}_m$) on the shape of production technology. Moving forward, we require the following assumption:

Assumption 6. $E[\boldsymbol{W}_i^\top \log(\epsilon_i)|\theta_{pi}, \theta_{mi}] = \mathbf{0}$

There remain two final challenges to be addressed. First, since the empirical model of time allocation can only infer unique values of $(\theta_{pi}, \theta_{mi})$ for students who chose minimal output $A_i \geq 2$ on our website, we have a missing regressors problem in equation (11). At this point, this is actually a fairly straightforward challenge to overcome: using the Tobit maximum likelihood results from the previous section, for each student i with A_i<2 we can compute the conditional expectations,[33]

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

The second challenge is that since student traits play the role of regressors in equation (11), sampling variability induces an errors-in-variables problem. We compute Empirical Bayes (EB) estimates of (θ_p, θ_m) in order to reduce attenuation bias by shrinking each fixed effect toward the mean in proportion to the individual noise in each 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, Pathak, Schellenberg, & Walters, 2020) is to assume a normal prior over the true fixed effect, $\log(\theta_{ji})$, and the estimation residual, r_{ji} for $j = e, l$. This implies a shrinkage factor of $\lambda_{ji} = \nu_j^2 \big/ \big(\nu_j^2 + \nu_{rji}^2\big)$, 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 $j = e, l$.[34] This results in the following EB estimates for student characteristics to be used as regressors for estimation of skill production technology:

$$\log(\theta_{pi})_{pB}{=}\lambda_{pi}\widehat{\log(\theta_{pi})}{+}(1{-}\lambda_{pi})\frac{\sum_{i=1}^N \widehat{\log(\theta_{pi})}}{N} \quad \text{and} \quad \log(\theta_{mi})_{pB}{=}\lambda_{mi}\widehat{\log(\theta_{mi})}{+}(1{-}\lambda_{mi})\frac{\sum_{i=1}^N \widehat{\log(\theta_{mi})}}{N}.$$

Finally, the un-balanced-panel nature of our data suggests that the error terms in equation (11) may exhibit heteroskedasticity. We formally test for this and find that the null hypothesis of homoskedastic errors is strongly rejected. Therefore, we estimate the production parameters via feasible generalized least squares in the familiar way, as outlined in Wooldridge (2016).

[33] This approach follows standard methods for regression with missing X's, surveyed by Little (1992, Section 4.2).

[34] 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}. Bootstrap estimation of ν_j^2 and ν_{rji}^2 are discussed in Section B.3.2.

Table 5. INITIAL MATH PROFICIENCY (Cobb-Douglas)

SPECIFICATION: **DEP. VARIABLE:** $\log(S_1)$	**(1)** (Mean; StDev)	**(2)** (Mean; StDev)	**(3)** (Mean; StDev)	**(4)** (Mean; StDev)
TFP ($\widehat{\log(A_i)}$)	(3.230; 0)	(3.105; 0.184)	(3.063; 0.208)	(3.071; 0.202)
θ_p **Prod. Share** ($\widehat{\alpha}_{pi}$)	(−0.346; 0)	(−0.289; 0.124)	(−0.258; 0.116)	(−0.263; 0.114)
θ_m **Prod. Share** ($\widehat{\alpha}_{mi}$)	(−0.017; 0)	(−0.019; 0.008)	(−0.025; 0.016)	(−0.024; 0.018)
	Mean StDev Effect	Mean StDev Effect	Mean StDev Effect	Mean StDev Effect
log(*TFP*)	N/A	0.4376***	0.4952***	0.4805***
(joint p-value)		($< 10^{-16}$)	($< 10^{-16}$)	($< 10^{-16}$)
log(θ_p)	-0.6711***	-0.5590***	-0.4991***	-0.5087***
(joint p-value)	($< 10^{-16}$)	($< 10^{-16}$)	($< 10^{-16}$)	($< 10^{-16}$)
log(θ_m)	-0.0625***	-0.0723***	-0.0926***	-0.0892***
(joint p-value)	(*0.0007*)	(4.2×10^{-5})	(1.1×10^{-7})	(*0.0002*)
CONTROL VARIABLES:				
District 2 ($\widehat{\alpha}_{01}, \widehat{\alpha}_{p1}, \widehat{\alpha}_{m1}$)	—	-0.2488***	-0.2666***	-0.2619***
(joint p-value)		($< 10^{-16}$)	($< 10^{-16}$)	($< 10^{-16}$)
District 3 ($\widehat{\alpha}_{02}, \widehat{\alpha}_{p2}, \widehat{\alpha}_{m2}$)	—	-0.6003***	-0.6816***	-0.6000***
(joint p-value)		($< 10^{-16}$)	($< 10^{-16}$)	($< 10^{-16}$)
Grade 5 ($\widehat{\alpha}_{03}, \widehat{\alpha}_{p3}, \widehat{\alpha}_{m3}$)	—	—	-0.2233***	-0.2126***
(joint p-value)			(1.9×10^{-10})	(1.4×10^{-9})
Female ($\widehat{\alpha}_{04}, \widehat{\alpha}_{p4}, \widehat{\alpha}_{m4}$)	—	—	-0.0450***	-0.0642***
(joint p-value)			(*0.0001*)	(*0.0007*)
Black ($\widehat{\alpha}_{05}, \widehat{\alpha}_{p5}, \widehat{\alpha}_{m5}$)	—	—	-0.1371***	-0.0953**
(joint p-value)			*(0.0042)*	*(0.0110)*
Hispanic ($\widehat{\alpha}_{06}, \widehat{\alpha}_{p6}, \widehat{\alpha}_{m6}$)	—	—	0.0428**	0.0452**
(joint p-value)			*(0.0272)*	*(0.0190)*
log(Mean Nbhd Income)	no	no	no	YES
Nbhd Uninsured Minor Rate	no	no	no	YES
# Peer & Adult Helper	no	no	no	YES
N	1,676	1,676	1,676	1,676
$\boldsymbol{R^2}$	0.423	0.492	0.518	0.521
Adjusted $\boldsymbol{R^2}$	0.422	0.490	0.513	0.513

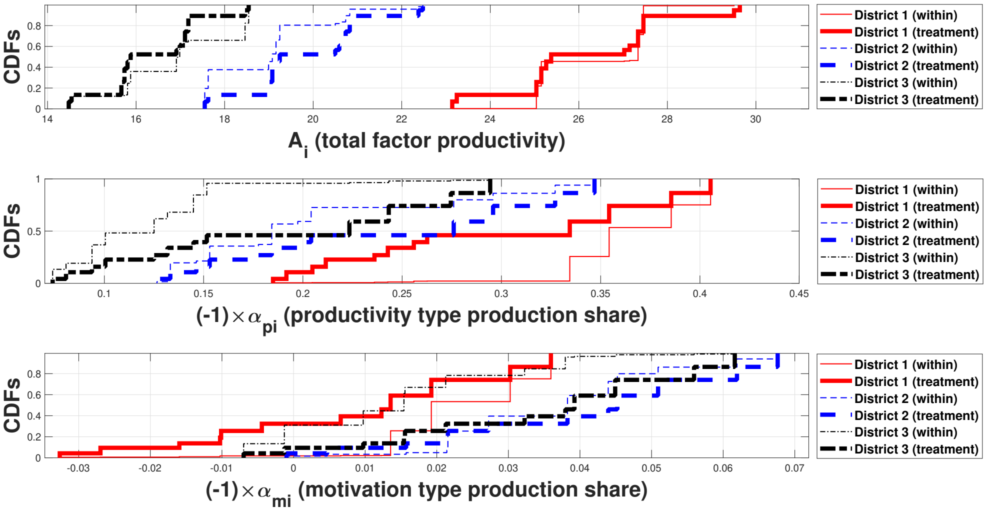
Notes: **Mean StDev Effect** is the total impact of a variable through both TFP (direct effect) and production shares of student inputs (interactions). For discrete variables **Mean StDev Effect** is the mean impact (across all students) of switching value from 0 to 1 (all else fixed), in standard deviation units of $log(S_1)$. For a continuous variable **Mean St. Dev. Effect** is the mean impact (across all students) of a one standard deviation increase (all else fixed), in standard deviations of $log(S_1)$. 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 three stars, two stars, and one star, respectively. In specification **(4)**, the interaction terms alone (i.e., $(\widehat{\alpha}_{pk}, \widehat{\alpha}_{mk})$, k=1,. . . ,6) have the following joint p-values: 1.4×10^{-7} for **District 2**; $<10^{-16}$ for **District 3**; 0.3314 for **Grade 5**; 0.0096 for **Female**; 0.0234 for **Black**; and 0.0086 for **Hispanic**. The p-value for a joint exclusion of the neighborhood socioeconomic terms and helper terms are 0.6224.

6.1.2. *Empirical Results.* Empirical results regarding the determinants of initial proficiency scores, S_i, are presented in Table 5. For ease of interpretation, rather than reporting coefficient values the table reports *standard deviation effects*, defined as the mean size (averaged across all students i) of a shift in $\log(S_1)$ that is induced (in standard deviation units of $\log(S_1)$) by an increase in a control variable of one standard deviation for continuous controls, or a 0-to-1 change for binary controls. These standard deviation effects encapsulate influence through all channels, both direct and indirect, but the lower caption of the table provides additional information to separate out effects on slopes.

Table 5 provides several interesting insights. First, we find that both θ_p and θ_m are significant determinants of initial math skill, but θ_p plays a clearly dominant role between the two. This insight should be considered alongside our earlier findings that females and Black students may be considered more motivated compared to other groups, having relatively more advantageous levels of θ_m, on average. Together, these results suggest new insights on educational interventions that aim to decrease gender or racial performance gaps in mathematics by motivating students through incentives or information about the returns to education (such as those studied in Fryer (2011); Levitt, List, and Sadoff (2016)).[35] These groups already tend

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

Figure 9. Idiosyncratic Cobb-Douglas Production Parameters by School District



Notes: Since θ_p (θ_m) is inversely related to productivity (motivation), the associated production share α_p (α_m) is *negative*. The lower two panels multiply production shares α_p and α_m by -1 for ease of interpretation; shifts to the right imply more productivity from a given factor. Thin lines represent CDFs for *students actually enrolled* in a given district, while thick lines represent general treatment effects, or model-implied CDFs for *all students* under enrollment at a given district.

to be more motivated than their male or White/Asian peers, suggesting that motivation is not the primary barrier limiting their progress. Moreover, (in specification (4)) since TFP is about 5.4 times as important as θ_m, and θ_p is 5.7 times as important, efforts to further incentivize marginal groups (further decreasing θ_m) may struggle to overcome the relative disadvantages these groups face.[36] We explore these considerations in more detail through counterfactual analyses in Section 6.3.

Second, we find strong evidence that school quality influences the production technology in important ways.[37] 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 productive human capital technology. Furthermore, the nature of the differences across school districts

[36] These insights may help explain why conditional cash transfers to students or families for increases in academic performance have often resulted in limited returns to learning (e.g., Fryer, 2011). Similarly, Levitt, List, and Sadoff (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. Levitt, List, Neckermann, and Sadoff (2016) show that incentives are more effective when delivered immediately. Cotton, Nordstrom, Nanowski, and Richert (2020) estimate returns from an intervention in developing countries providing girls, their families, and communities with information about the benefits of girls' education, while motivating the academic efforts of the girls. They find that such interventions can have significant effects on academic progress, but at potentially prohibitive costs.

[37] In discussing quality, it is important to note that the mechanisms through which school value added arise are not well understood. For example, the research team's interviews with teachers and administrators in District 3 (the low performing district) suggest a high level of engagement, commitment, and effort among faculty and administrators. One principal claimed to know all of his 600+ students by name, and provided ample evidence to that effect during a visit which lasted several hours. At the same time, in his/her school, where resources to employ full janitorial staff were lacking, faculty would take turns cleaning the cafeteria room during lunch periods. More research is needed to understand how internal resource constraints and institutional structure are tied to school value added.

is not merely one of *levels*, but of the fundamental *shapes* of the production processes employed. Figure 9, which plots empirical CDFs of student-specific production parameters, illustrates an interesting and novel finding: high-performing school districts have higher TFP and lean more heavily on study-time productivity, whereas middle- and low-performing schools have lower TFP and lean more heavily on a student's motivation level to generate improvements in math skill.[38]

Third, we also find evidence of decreasing returns to scale production technology in the sense that $-(\alpha_p+\alpha_m)$ is well below a value of 1 (which would indicate constant returns to scale) 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 and more favorable. 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. When modeling short-run proficiency gains during our sample period, we can go one step further by incorporating available data on interim extracurricular math learning activity. The model allows student study effort to improve test performance through either total study time, T_i, volume of completed learning tasks, A_i, or both. Formally, we model incremental proficiency gains as a flexible quadratic polynomial, where once again the shape of the production technology is idiosyncratic to child i and depend on her pre-existing characteristics:

$$\Delta S_i = \Delta_{0i} + \Delta_{1i}T_i + \Delta_{2i}T_i^2 + \Delta_{3i}A_i + \Delta_{4i}A_i^2 + \Delta_{5i}(T_i \times A_i) + \varepsilon_i. \tag{12}$$

Once again, ε_i is an idiosyncratic, transitory shock. Regression parameters are once again sub-scripted by i because they depend on individual student covariates, with $\Delta_{ji} \equiv \boldsymbol{V}_i\boldsymbol{\delta}_j$ for $j = 0, 1, ..., 5$, being a single index of covariate vector $\boldsymbol{V}_i = [\boldsymbol{W}_i, S_i, \theta_{pi}, \theta_{mi}]$, which encompasses the full set of controls from the previous section, and includes initial proficiency and student traits as additional controls.[39]

By including the structural student types $(\theta_{pi}, \theta_{mi})$ in $\boldsymbol{V}_i$, we allow them to play a dual role in shaping a student's ability to acquire new skill: first, they underlay choices of T_i and A_i, and second, they may alter the rate at which a fixed volume of study activity is converted into measurable proficiency gains. Including initial proficiency S_i as a control allows for possible decreasing-returns-to-scale technology where incremental gains of a fixed size become more difficult as a student achieves greater subject mastery. Finally, note that our model of incremental proficiency gains allows for school quality, contained in $\boldsymbol{W}_i$, to impact learning through 3 distinct channels: (*i*) it can influence a child's productivity and motivation level (through equations 7); (*ii*) it can directly impact learning independently of at-home math activity (through influencing the intercept term $\Delta_{0i} = [\boldsymbol{W}_i, S_i, \theta_{pi}, \theta_{mi}]\boldsymbol{\delta}_0$), and (*iii*) it can alter the shape of the mapping between (T_i, A_i) and ΔS_i (through influencing the coefficient terms $\{\Delta_{1i}, \dots, \Delta_{5i}\}$). It will be shown hereafter that all three channels of school-quality influence are present and economically meaningful.

[38] While this result is novel within the education literature, it has interesting parallels to the literature on production technology estimation in industrial organization, where it has been shown that firms with access to systematically different inputs often evolve their productive technologies accordingly.

[39] For numerical stability in our short-run production function analysis, we normalize T (practice time in minutes) and initial test score S_1 by subtracting means and dividing by standard deviation.

6.2.1. *Estimating the model.* We require the following assumption on short-run proficiency gains shocks:

Assumption 7. $E[\boldsymbol{V}_i^\top \varepsilon_i | \theta_{pi}, \theta_{mi}] = \mathbf{0}$.

The empirical strategy here faces similar challenges as in the previous Section 6.1, and we therefore employ similar coping strategies, including empirical Bayes shrunk forecasts and feasible GLS estimation. The results of this analysis are presented in Table 6. We again summarize results as *standard deviation effects* rather than reporting long lists of (up to 78) parameter estimates, though an adjustment is in order. In regression analysis standard deviations are commonly used as units of "typical" shift for a random variable, but they lose that intuitive meaning as the distribution becomes more skewed.[40] Such is the case for T and A (see Table 1, Figure 2), where standard deviations exceed the respective 80th percentiles. The usual standard deviation would constitute an especially extreme hypothetical shift in behavior for the 50% of students who did no work on the website. Thus, we define *pseudo-standard deviation* (pStDev) as $pStDev_j \equiv F_j^{-1}(0.5|worker) - F_j^{-1}(0.159|worker)$, $j = t, a$, for computing standard deviation effects. The pStDev is defined this way because for normally-distributed data it reduces to the usual standard deviation, and it provides a more meaningful measure of a "typical" unit of shift for the average child in the sample. Pseudo-standard deviations for T and A (relative to all students, not just workers) are roughly 64 minutes of focused problem solving time and 8.4 website tasks completed (i.e., 50.4 practice problems solved).

6.2.2. *Empirical Results.* In Table 6 we find that learning task completion A_i (and not simply time spent studying) is primarily responsible for short-term gains in adolescent math skill. Indeed, StDev Effect estimates for T_i are consistently *negative*, meaning that (holding all else fixed, including productivity types θ_p) total time spent actually plays the role of tempering (but never swamping) the conversion rate of task completion A_i into short-term gains in measured math proficiency. While the StDev Effects for T_i and A_i seem large, one should recognize that it is not a well-posed thought experiment to hold one fixed while varying the other, like it is for other regressors. This is because time spent working achieves output volume with positive probability, and work tasks cannot be completed without time inputs. A more meaningful interpretation would involve a simultaneous pStDev increase in both T_i and A_i, which on average would entail a net increase of $1.375 = 2.152 - 0.777$ standard deviations of skill gain ΔS.

Table 6 provides further evidence of a decreasing returns to scale human capital production technology: the estimated StDev Effect of pre-test score S_i is significant (both statistically and economically) and *negative*. This implies that as students reach a higher level of mastery of math concepts, achieving further improvements of a fixed size (in test score space) becomes more and more difficult. On the other hand, all else equal, the results suggest that as children progress from 5th-grade to 6th-grade they become more effective at learning, with the year-on-year difference being equivalent to roughly one quarter SD of skill gains ΔS. 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 find that θ_p also alters the shape of the short-run learning technology in an economically meaningful way. That is, students who progess through learning tasks more quickly also tend to derive more incremental permanent skill from those tasks as well. This effect comes both directly through the intercept, and indirectly through the slope terms. Finally, we find once again that after controlling for the rich set of

[40] As an extreme but illustrative counterexample, one would hesitate to interpret standard deviation as a typical unit of shift for a Pareto-distributed random variable, which may exhibit large or infinite variance due to a small mass of extreme values.

Table 6. PRODUCTION OF INCREMENTAL GAINS IN MATH SKILL

SPECIFICATION:	**(1)**	**(2)**	**(3)**	**(4)**
DEP. VARIABLE: ΔS	Mean StDev Effect	Mean StDev Effect	Mean StDev Effect	Mean StDev Effect
$\boldsymbol{T}$ **(standardized)**† $\left(\widehat{\Delta}_{1i}, \widehat{\Delta}_{2i}, \widehat{\Delta}_{5i}\right)$	0.1477***	-0.5464***	-0.6183***	-0.7770***
(joint p-value)	*(4.1×10^{-5})*	*($< 10^{-16}$)*	*($< 10^{-16}$)*	*($< 10^{-16}$)*
$\boldsymbol{A}^{\dagger}$ $\left(\widehat{\Delta}_{3i}, \widehat{\Delta}_{4i}, \widehat{\Delta}_{5i}\right)$	0.3196***	1.4238***	1.6101***	2.1520***
(joint p-value)	*(0.0013)*	*($< 10^{-16}$)*	*($< 10^{-16}$)*	*($< 10^{-16}$)*
$\boldsymbol{S_1}$ **(standardized)** $(\widehat{\delta}_{0,1}, \ldots, \widehat{\delta}_{5,1})$	—	-0.4169***	-0.4334***	-0.4380***
(joint p-value)		*($< 10^{-16}$)*	*($< 10^{-16}$)*	*($< 10^{-16}$)*
$\mathbf{log}(\boldsymbol{\theta_p})$ $(\widehat{\delta}_{0,2}, \ldots, \widehat{\delta}_{5,2})$	—	-0.3446***	-0.3019**	-0.3233***
(joint p-value)		*(0.0040)*	*(0.0119)*	*(3.3×10^{-5})*
$\mathbf{log}(\boldsymbol{\theta_m})$ $(\widehat{\delta}_{0,3}, \ldots, \widehat{\delta}_{5,3})$	—	-0.0231**	0.0364**	0.0542***
(joint p-value)		*(0.0370)*	*(0.0233)*	*(0.0008)*
District 2 $(\widehat{\delta}_{0,4}, \ldots, \widehat{\delta}_{5,4})$	—	-0.1593***	-0.2277***	-0.1261***
(joint p-value)		*(0.0002)*	*(4.1×10^{-8})*	*(9.3×10^{-8})*
District 3 $(\widehat{\delta}_{0,5}, \ldots, \widehat{\delta}_{5,5})$	—	-0.4295***	-0.5765***	-0.4435***
(joint p-value)		*($< 10^{-16}$)*	*($< 10^{-16}$)*	*(2.2×10^{-6})*
Grade 5 $(\widehat{\delta}_{0,6}, \ldots, \widehat{\delta}_{5,6})$	—	—	-0.2323***	-0.2256***
(joint p-value)			*(5.3×10^{-5})*	*(2.1×10^{-8})*
Female $(\widehat{\delta}_{0,7}, \ldots, \widehat{\delta}_{5,7})$	—	—	0.0269	0.1257**
(joint p-value)			*(0.4085)*	*(0.0164)*
Black $(\widehat{\delta}_{0,8}, \ldots, \widehat{\delta}_{5,8})$	—	—	0.0765***	0.1118***
(joint p-value)			*(2.0×10^{-8})*	*(2.2×10^{-7})*
Hispanic $(\widehat{\delta}_{0,9}, \ldots, \widehat{\delta}_{5,9})$	—	—	0.0704***	0.0900**
(joint p-value)			*(0.0018)*	*(0.0311)*
#Tot. Regressors (incl. interactions)	7	37	61	79
log(Mean Nbhd Income)	no	no	no	YES***
Nbhd Uninsured Minor Rate	no	no	no	YES**
# Peer & Adult Helper	no	no	no	YES***
$\boldsymbol{N}$	1,494	1,494	1,494	1,494
$\boldsymbol{R^2}$	0.096	0.196	0.220	0.229
Adjusted $\boldsymbol{R^2}$	0.093	0.177	0.188	0.187

Notes: For context, the mean 2-week learning, ΔS, by marginal/inactive students (who did no extracurricular math on the website) was 0.51, with 95% confidence interval [0.28,0.74]. **Mean StDev Effect** is the total impact of a variable through the intercept Δ_{0i} (direct effect) and slope terms $(\Delta_{1i}, \ldots, \Delta_{5i})$ (interactions). For discrete variables **Mean StDev Effect** is the mean impact (across all students) of switching from 0 to 1 (all else fixed), in standard deviation units of ΔS. For a continuous variable **Mean St. Dev. Effect** is the mean impact (across all students) of a one standard deviation increase (all else fixed), in standard deviations of ΔS.

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 three stars, two stars, and one star, respectively. In specification **(4)**, the interaction terms alone (i.e., $(\widehat{\delta}_{1k}, \ldots, \widehat{\delta}_{5k})$, k=1,...,9) have the following *joint p-values*: 0.0004 for $\boldsymbol{S_1}$ (standardized pre-test score); 0.0008 for $\mathbf{log}(\boldsymbol{\theta_p})$; 0.0005 for $\mathbf{log}(\boldsymbol{\theta_m})$; 9.8×10^{-8} for **District 2**; 9.3×10^{-6} for **District 3**; 2.2×10^{-6} for **Grade 5**; 0.0889 for **Female**; 3.3×10^{-6} for **Black**; and 0.1138 for **Hispanic**. Neighborhood socioeconomic proxies are statistically significant (*joint p-values* of *(0.0008)* and *(0.0452)*, respectively) but collectively play a minor role in predicting proficiency changes.

†Due to heavily skewed distributions of T and A, rather than using their standard deviations to compute **Mean St. Dev. Effect**, we use the *pseudo-standard deviation*, (defined above) instead. For normally distributed data, pStDev=standard deviation.

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 on average reduces skill augmentation by 0.13 SD and 0.44 SD, respectively. Once again, our confidence in attaching causal interpretations to these results is bolstered by the stability of estimated magnitudes across model specifications that include school assignment as a control. This is suggestive of having adequately controlled for relevant sources of omitted variable bias with our large set of additional controls and interactions.

In interpreting the results from Table 6 regarding StDev effects, one should keep in mind 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 school districts, different initial proficiencies, different unobserved traits, different home backgrounds different genders, and different racial backgrounds. In the following section we highlight rational study choices and learning effects when varying one key component of adolescent life circumstances: school assignment.

6.3. **Model Simulations and Counterfactual Analysis.** We can now execute counterfactual experiments to investigate the role of access to high-quality education services in explaining racial achievement gaps within our sample population. For Black and Hispanic students, the profile of schools attended is heavily tilted toward middle- and low-performing schools and away from the highest-performing school district. Holding school assignment fixed for White/Asian students, we alter school assignment for Blacks and Hispanics by repeatedly re-sampling (with replacement) from the distribution of school assignment among Whites and Asians. Intuitively, this exercise levels the playing field by bringing Black/Hispanic school quality allocation *up* to the empirical level of White/Asian school assignment, while leaving the latter fixed.[41] We then use model estimates to compute adjusted θ_p^* under the new school assignments, and we simulate counterfactual distributions of pre-exam scores and choices of T and A under our existing incentive schemes. For each minority student we re-simulate counterfactual school assignment many times to wash out the role of simulation error in driving our results.

In the previous sections we saw evidence for three distinct roles of school quality in driving academic choices and outcomes: (i) influencing productivity and motivation, (ii) directly augmenting a child's learning independent of home-study activity, and (iii) enhancing the conversion of fixed quantities of learning activity into incremental skill gains. There remains a question of what, exactly, is entailed in the exercise of bringing Black/Hispanic school quality endowment up to a similar level enjoyed by Whites/Asians. Better facilities? More effective administrators and teachers? Better educational support resources? Better peer effects? Identifying specific channels through which school quality effects are manifest is beyond the scope of this paper, but it is likely a mixture of these. To the extent that one has a good idea of which channels are most salient for delivering school-quality impact, one can interpret our counterfactual school resource equalization as a prescription for a specific set of interventions. Otherwise, our counterfactuals are indicative of the role played holistically by inequities of publicly-provided educational inputs in driving racial achievement gaps, holding an extensive list of other controlled factors of a child's life fixed, including structurally estimated idiosyncratic productivity and motivation types (η_{pi}, η_{mi}), gender, race, affluence, home resources, family academic support, consumption proxies, parental permissiveness for electronic entertainment, academic attitudes, intrinsic/extrinsic mindedness, and a host of time-use factors like screen time, peer social time, enrollment in sports, enrollment in music programs, and participation in organized clubs.

6.3.1. *Racial Achievement Gaps.* The model predicts complex changes to racial achievement gaps that vary by a child's percentile rank within her demographic group. These are depicted graphically in Figures 10 and 11, and numerically in Table 7. Generally, the closure of the racial achievement gaps from academic resource equalization becomes more pronounced among higher achieving students. Indeed, our model predicts that bringing Black/Hispanic school quality up to the same level as empirically exists for Whites/Asians would

[41] An alternative exercise would be to simply re-allocate all existing school seats via a lottery. Both methods would hypothetically level the playing field, though the one we adopted—interpretable as a new infusion of resources targeted at the Black/Hispanic communities—doesn't require grappling with re-distribution concerns and also has an interesting interpretation in terms of implications for affirmative action in college admissions.

Figure 10. Counterfactual Achievement Gaps: Black vs White/Asian

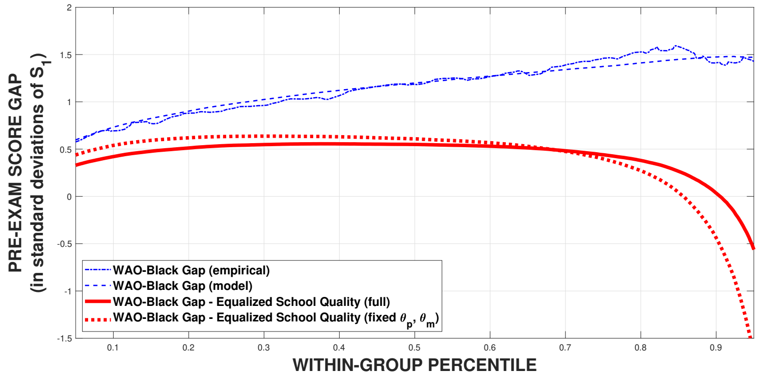
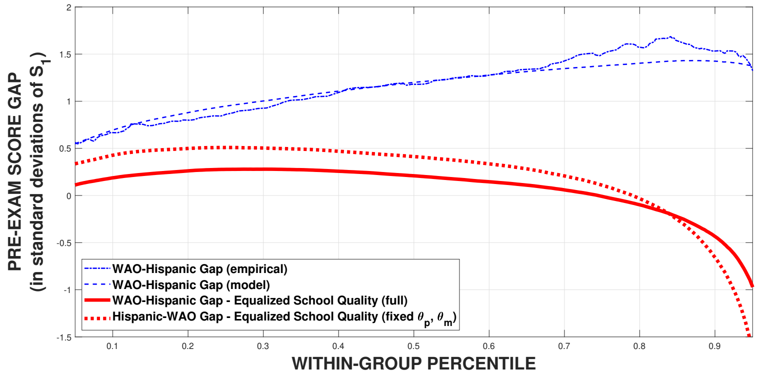


Figure 11. Counterfactual Achievement Gaps: Hispanic vs White/Asian



Notes: for $r \in (0.05, 0.95)$ Figure 10 (Figure 11) depicts the empirical and counterfactual differences in exam scores between a child at the r^{th} percentile within the White/Asian group and a child at the r^{th} percentile within the Black (Hispanic) group.

Table 7. SCHOOL-QUALITY EQUALIZATION: ACHIEVEMENT GAPS

	PERCENT CHANGE IN ACHIEVEMENT GAPS AT:					
	10th Pctl	**25th Pctl**	**Median**	**75th Pctl**	**90th Pctl**	**Mean Integrated % Change**
Black *(full schl. qual. equalization)*	−42.5%	−44.6%	−54.2%	−68.0%	−97.6%	−53.5%
Black *(fixed (θ_p, θ_m))*	−26.3%	−34.7%	−49.3%	−71.1%	−129.7%	−52.9%
Hispanic Students *(full schl. qual. equalization)*	−73.0%	−70.8%	−82.6%	−100.7%	−130.5%	−80.2%
Hispanic Students *(fixed (θ_p, θ_m))*	−38.5%	−46.3%	−65.6%	−92.0%	−146.0%	−67.0%

cause the highest performing Black and Hispanic students to actually overtake their White/Asian counterparts in terms of exam score performance. Integrating over gap closure magnitudes at different percentiles generates a single aggregate summary value: holding all other student characteristics fixed, racial differences in school quality account for roughly 54% of the achievement gap between Blacks and Whites/Asians in our sample, and roughly 80% of the achievement gap between Hispanics and Whites/Asians. We also ran an alternate specification of this counterfactual achievement gap calculation, where we held underlying θ_p

fixed, and only vary the production technology with the counterfactual school assignment profile. This decomposition reveals that most of the achievement gap narrowing for Blacks and Hispanics is due to changes in the long-run production technology that exist at higher-quality schools, holding student traits fixed.

6.3.2. *Using Affirmative Action to Offset School Quality Differences in Academic Contests.* Building on the results of the previous exercise, we also consider a hypothetical head-to-head academic competition between all students in our sample. This hypothetical competition assumes a large-market, many-to-many, contest structure familiar to college admissions models in Bodoh-Creed and Hickman (2018), and Cotton et al. (2022); Cotton, Hickman, and Price (2020), in which students compete for admissions to an array of vertically-differentiated universities by investing in their observable human capital (as measured by grades/test scores). We use the simulation results from the first counterfactual to ask, "What would the Affirmative Action scheme have to be in order to exactly wipe out the ex-ante advantage to White/Asian students which comes not from having better household or individual characteristics, but from simply attending better schools?"

Intuitively, in rank-order contests like college admissions, there may exist systemic, arbitrary disadvantages to some competitors before the competitive human capital investment game begins. Using our results, we can quantify the precise affirmative action scheme that would ex-post remove that systemic disadvantage, and nothing more. The results of this calculation are displayed in Figure 17. For this exercise we combine Blacks and Hispanics into a single, composite, underrepresented minority group for simplicity. The horizontal axis displays URM percentiles, and the vertical axis is a point-specific score bonus (in standard deviation units of the original pre-test scores). For comparison, the plot also depicts a baseline rule, commonly referred to as "color-blind" admissions, which is simply a constant zero-bonus for all minority students.[42] Note that the plot zooms in on the 5-95 range since behavior in the extreme tails fodfxr model simulations can be less reliable. The salient features of the equal-school-equivalent AA scheme are (I) the score bonus is substantially above the race-blind alternative along the entire distribution of URM students; and (II) it trends steadily upward for the highest achievers. This novel result based on our causal estimates of student characteristics and value-added estimates of school inputs may have important implications for the ongoing legal debate surrounding affirmative action in college admissions.

6.3.3. *Incentive response counterfactuals.* Finally, we seek to better understand the extent to which a policymaker could lean on the incentive channel alone to close achievement gaps by inducing Black and Hispanic students to increase math activity. We also ran a similar analysis to see how hypothetical school quality equalization would impact the answer to this question. The general take-home lesson from this section is that, without getting more serious about equalizing the quality of public education inputs accessible to Black and Hispanic students, the incentive lever does not appear as a terribly promising option for a policymaker.

More concretely, Figures 12 and 13 explore what we refer to as *Incentive Response Gaps.* To define that term, first note that an *Incentive Response Function* (IRF) is defined as the difference in the quantile functions of A (or T alternatively) under different contracts. For example, the White/Asian *Incentive*

[42] It is worth mentioning that the results in this section call into question the appropriateness of the common label "color-blind admissions" for the baseline rule, given that it ignores a large asymmetry of causal value-added resources delineated by a child's race. We maintain the common label here simply for its familiarity within the public debate on affirmative action.

Figure 12. Incentive Response Gaps in Learning Activity: Black vs White/Asian

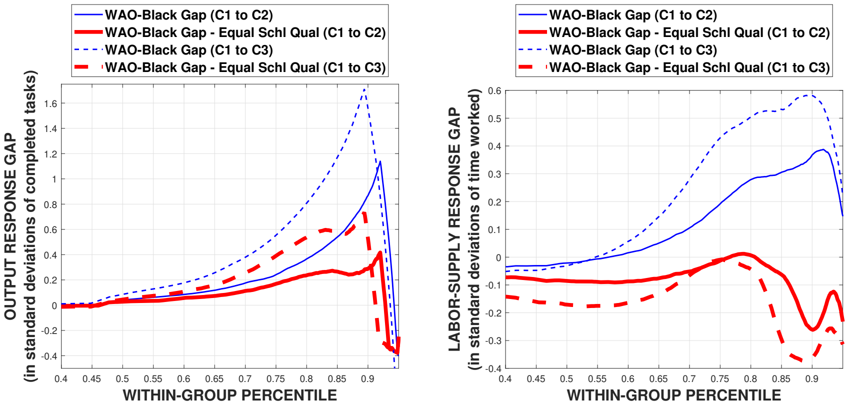
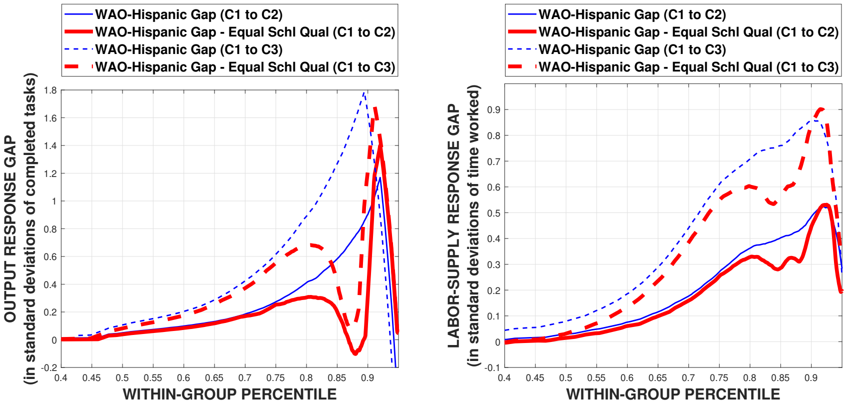


Figure 13. Incentive Response Gaps in Learning Activity: Hispanic vs White/Asian



Notes: Incentive Response Gaps depict differences across race groups in marginal learning activities under strengthening of incentives from contract 1 to contract 2 or contract 3. For each $r \in (0.05, 0.95)$, the Figure 12 (Figure 13) depicts the difference between increased output for a student at the r^{th} percentile within the White/Asian group, and a student at the r^{th} percentile within the Black (Hispanic) group. Thin lines depict IRGs under the status quo and thick lines represent IRGs under the school-quality equalization counterfactual.

Response Function for a contract 1-to-contract 2 shift would be

$$IRF(j, W/A, 1, 2) \equiv F_j^{-1}(r|W/A, contract\ 2) - F_j^{-1}(r|W/A, contract\ 1),\ j = q, t,\ r \in [0, 1], \tag{13}$$

or the quantile function of A or T for Whites/Asians under contract 2, minus the corresponding quantile function for Whites/Asians under contract 1. This measures, at various percentiles of the student distribution, how students respond to an increase in piece-rate incentives. With that definition in mind, the Black-White/Asian *Incentive Response Gap* (IRG) is the IRF for Whites/Asians under a contract 1-to-contract 2 shift, minus the IRF for Black students under the same contract 1-to-contract 2 shift. The IRG therefore measures the *difference* across race groups in their responsiveness to piece-rate incentives. For example, if $IRG(0.5|j, Black, White/Asian, 1, 2) = 5$, that would mean that when the median White/Asian student is switched from contract 1 to contract 2, she increases her total output on dimension $j = q, t$ by 5 units *more* than the median Black student under the same shift in incentives.

From our earlier analysis, one might believe that since Black students have systematically higher motivation (lower θ_m), that they would be more responsive to incentives. However, such intuition is incomplete, and it is important to recognize that one's study effort is determined by the interaction between a student's motivation *and* how much time is needed for task completion, which is a function of θ_p. While it is true that a lower θ_m makes it less burdensome for a student to give up an hour of would-be leisure time, higher values of θ_p work in the opposite direction and make a student's time less valuable for earning rewards of time spent working. Moreover, due to the dramatic curvature in the utility cost function, it turns out that θ_p is quite crucial for inducing students to respond to incentives and increase learning task accomplishment.

With these ideas in mind, Figures 12 and 13 plot the IRGs under the status-quo and under school quality equalization. The left panels shows activity output A and the right panels show time worked T. Incentive responses and response gaps are fairly low until the 75th percentile (i.e., most studious) students. In that upper region the response gaps in terms of A are quite substantial, but are reduced significantly by equalizing school quality, with its implied increase of study-time productivity (i.e., reduction in θ_p). Note also that the incentive response gaps are smaller in terms of T, and also change less in terms of T. This reflects the fact that because of the huge curvature of the utility cost function $c(t;\widehat{\gamma}_c)$, learning gains under optimal labor-leisure choice are primarily accomplished through increases in the productivity of time, rather than through large re-allocations of a child's time from leisure toward math.

Figures 14 and 15 consider a somewhat more drastic experimentation with piece-rate incentives. On the horizontal axis are different simulated contract offerings, this time with no lump-sum base wages for simplicity. Once again, the left panels plot simulated activity output and the right panels plot labor supply. Thin lines represent the status-quo school assignment and thick lines represent the re-sampled, equalized, school quality regime. Each of the plots in Figures 14 and 15 depict the behavior of the median most studious student, and the 25th (less studious) and 75th (more studious) percentiles for all students, including both workers and non-workers in the experimental data. These figures provide the clearest illustration of why the incentive channel is relatively weak. For example, in order to induce the 75th percentile most studious Hispanic student (Figure 15) to produce roughly 12 units of learning-by-doing tasks (under status-quo school assignment) the policy-maker would have to offer an outlandishly high piece rate of \$16 per quiz.

To be clear, θ_m *does* matter: the 75th percentile most studious Black student (Figure 14) would produce about 35 units of learning-by-doing tasks at \$16 per quiz, and the biggest difference between the two groups is the distribution of θ_m. However, for both groups overcoming their disadvantage in terms of θ_p through the incentive channel alone requires very large financial incentives. Now, consider a comparison of this outcome for the status quo setting, in which the current distribution of students across school districts is held constant, to the outcomes from a counterfactual setting in which minority groups have identical access to school quality as Whites/Asians. For minority students, such a shift in school district produces large improvements in study-time productivity θ_p while leaving θ_m largely untouched. In such a scenario, under-served minority students become dramatically more responsive to piece-rate incentives (thick lines), as depicted in Figures 14 and 15.

Figure 14. Incentive Response in Learning Activities: Black

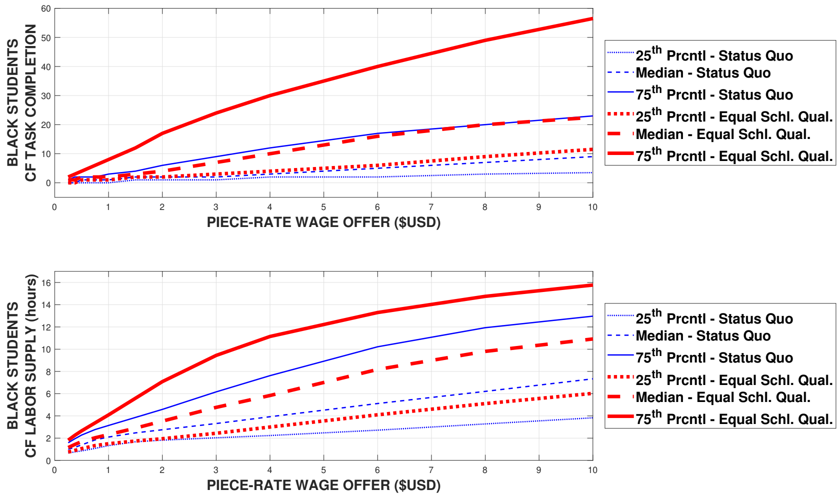
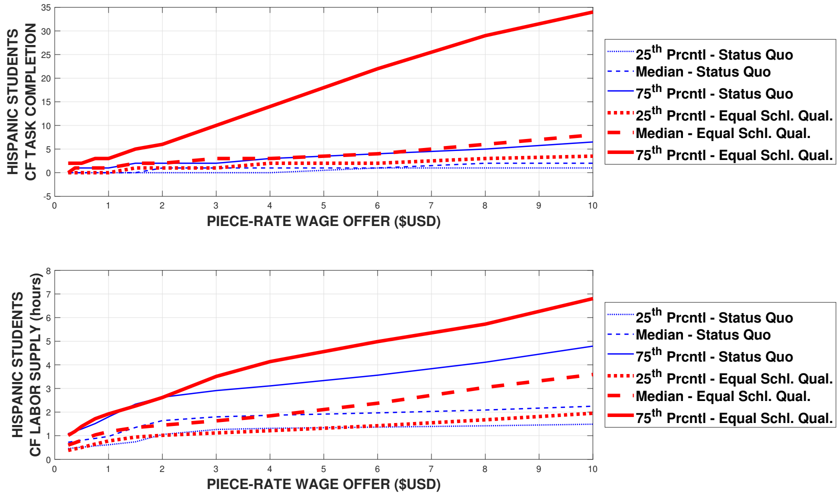


Figure 15. Incentive Response in Learning Activities: Hispanic



7. Conclusion

There are several important lessons for education policy to come out of our analysis. 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 more than motivating under-performing students. We find that large fractions of struggling students are already willing to spend time studying. Rather, 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 differences in quality across school districts, suggesting that individual students enrolled in less-affluent districts are at a substantial disadvantage compared to students in higher performing districts. Students in less-affluent districts tend to be both less motivated and less productive with their study time, even after controlling for student observables and socioeconomic factors. Our analysis relies on a sophisticated structural model and econometric analysis to directly quantify motivation and study-time productivity parameters for each individual student. In so doing, we uncover evidence of a causal link between school quality and students being less willing to spend time studying, and less productive in their home studies. School quality not only affects the preferences and performance of individual students, it also contributes to the education performance gaps between racial groups. Through a counterfactual analysis of the structural model, we show that school district differences drive an estimated 54% (80%) of the current test score gap between Black (Hispanic) and White/Asian students in our sample. Such insights suggest that having access to better performing schools are likely to have significant impacts on one's learning process and academic achievement, and highlights the potential need to better target resources and educational support at under-performing districts to ensure that all students build foundational literacy and mathematics skills.

Since the 1960s one would be hard pressed to find two disciplines within economics that have grown more and established as many deep insights as the study of the role of human capital on economic growth and the study of how education, learning, and skills are produced. Likewise, a perusal of the popular press suggests that most have accepted James Mill's dictum that "if education cannot do everything, there is hardly anything it cannot do." Yet, even with these movements, modern economies continue to seek ways to increase the proportion 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 within an economy.

A lesson gleaned from the work of Heckman and colleagues, as well as many others, is that investment in human capital pays off at a greater rate than does investment in physical capital, which suggests that we must move from an economy of scarcity of educational opportunity to one of promoting and developing

[43] Such insights are consistent with several past evaluations, which find relatively small impacts on test scores from education interventions focused on information provision or student incentives for studying (e.g., Baird, McIntosh, and Ozler (2021)), or which find more substantial impacts from programs that involve either a strengthening of early grade foundational math and literacy skills (e.g., Banerjee et al. (2016)), or adapting curriculum and teaching to the individual needs of learners, whether though tutoring, formative assessment, individualized education plans, or technology assisted learning (e.g., Pitchford, Chigeda, and Hubber (2019), Outhwaite, Gulliford, and Pitchford (2017), Rodriguez-Segura (2020)). Cotton et al. (2021) conducts systematic cost-benefit analyses of alternative education programs in Malawi using impact evaluation data from various settings and concludes that a technology assisted learning program that enables "teaching at the right level" is the most cost effective means of improving education outcomes.

all students over the life-cycle. A troubling observation from our raw data that underscores the current state of developmental resource scarcity is that, while Black and Hispanic students in our sample self-report higher preferences for studying math and science relative to other academic subjects, they are vastly less affluent, much more likely to lack health insurance coverage, and are almost entirely relegated to schools with average or below-average instructional budgets, faculty salaries, and teacher degree qualifications. Their standardized test scores unsurprisingly lag far behind their White/Asian counterparts—slightly more than a full standard deviation in our math pre-test, on average—whose corresponding resource allocations on all the above dimensions are almost entirely at average or above-average levels, relative to the rest of the State of Illinois. These facts together suggest adults are successfully advertising to Black and Hispanic children that math and science education are the way out of poverty. However, their communities, schools, and society at large are failing to follow up on the marketing campaign by equipping them with the tools to effectively act upon this perception.

Of course, any particular exercise leaves much on the sidelines. In our case, we should be clear that we believe academic efficiency and time preference are not completely stable over the long run. There is ample evidence (Bloom, 1964; Hunt, 1961) that academic efficiency may 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 the teacher instruction (which could be due to poor 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, 3$^{\text{rd}}$ ed.* Chicago: University of Chicago Press.

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

Berger, J., & Pope, D. (2011). Can losing lead to winning? *Management Science*, *57(5)*, 817–827.

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. (2018). College assignment as a large contest. *Journal of Economic Theory*, *175*, 88–126.

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. (2020). Affirmative action, shifting competition, and human capital accumulation: A comparative static analysis of investment contests. *Queen's University working paper*.

Cotton, C., Hickman, B. R., & Price, J. P. (2022). Affirmative action and human capital investment: Evidence from a randomized field experiment. *Journal of Labor Economics*, *40(1)*, 157–185.

Cotton, C., Kashi, B., MacKinnon, J., Makuwira, J., Nordstrom, A., Wallace, L., ... Wong, B. (2021). Cost-benefit analysis: Improving the quality of primary school education in malawi. *Malawi Priorities Project*. (National Planning Commission of Malawi)

Cotton, C., Nordstrom, A., Nanowski, J., & Richert, E. (2020). Improving girls education outcomes through community-wide information and empowerment campaigns. *Queen's University 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., & Weidemann, C. (2023). 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* (2$^{\text{nd}}$ *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. (2019). Toward and understanding of corporate social responsibility: Theory and field experimental evidence. *NBER Working Paper No. 26222*.

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.

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., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, *8*, 183–219.

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)

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.

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.

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

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25(1)*, 68–81.

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$^{\text{th}}$graders and

16 were developed for 6thgraders.[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. All 672 problems were pooled to expose both 5thand 6thgraders 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 6thgrade materials and the other 16 from 5thgrade 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 $\approx$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]

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

[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 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 5thand 6thgrades (`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.

[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

Table 8. DESCRIPTIVE STATISTICS: STUDENT COVARIATES BY SUB-SAMPLE

SUB-SAMPLE:	**ALL**	**FEMALE**	**MALE**	**BLACK**	**HISPANIC**	**WHITE/ASIAN**
SIZE/FRACTION OF TOT.:	1,676	0.5078	0.4922	0.2691	0.1915	0.5394
SCHOOL DISTRICT & NEIGHBORHOOD SOCIOECONOMICS						
Nbhd Mean Income	$108,917	$108,917	$108,917	$80,774	$45,687	$132,038
(sample std. dev.)	*(41,470)*	*(41,107)*	*(41,871)*	(32,390)	*(23,175)*	*(24,602)*
Nbhd Uninsured Minors	0.252	0.253	0.252	0.378	0.616	0.072
(sample std. dev.)	*(0.297)*	*(0.297)*	*(0.297)*	*(0.293)*	*(0.231)*	*(0.129)*
District 1	0.465	0.475	0.455	0.007	0.044	0.843
District 2	0.268	0.260	0.276	0.650	0.103	0.136
District 3	0.267	0.266	0.269	0.344	0.854	0.021
FAMILY & RECREATIONAL TIME-USE VARIABLES						
# Adult Academic Helpers	1.140	1.163	1.117	1.128	0.615	1.328
(sample std. dev.)	*(0.848)*	*(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 std. dev.)	*(0.783)*	*(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 std. dev.)	*(1.135)*	*(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 std. dev.)	*(0.327)*	*(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 std. dev.)	*(1.201)*	*(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 std. dev.)	*(0.471)*	*(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 std. dev.)	*(0.487)*	*(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 std. dev.)	*(0.492)*	*(0.496)*	*(0.486)*	*(0.473)*	*(0.465)*	*(0.500)*
Fraction of Peer Social Time In Adult-Supervised Activities	0.351	0.356	0.345	0.317	0.274	0.392
(sample std. dev.)	*(0.172)*	*(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 std. dev.)	*(0.480)*	*(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 std. dev.)	*(0.411)*	*(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 std. dev.)	*(1)*	*(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 std. dev.)	*(1)*	*(1.005)*	*(0.992)*	*(1.047)*	*(1.057)*	*(0.949)*

Notes: Unless otherwise stated, standard font numbers in the table represent sample means, while italicized numbers in parentheses represent sample standard deviations. **Adult Academic Helpers** included parents, grandparents, and tutors; **Peer Academic Helpers** included siblings and friends. Numbers reported for Neighborhood Mean Income represent the median across all students in the sample. **Extrinsic Motivation Score** and **Intrinsic Motivation Score** both exist on a scale of 0-4, but have been standardized for this table. All other figures represent sample means, with sample standard deviations in parentheses and italics. Fifth-graders make up 47.3% of the total sample, with 6$^{\text{th}}$graders comprising the other 52.7%. Sub-sample proportions are close to that ratio for all gender and race groups.

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

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.

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.[48] Females in our sample also self-reported higher levels of intrinsic motivation, and lower levels of extrinsic motivation, relative to males.

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

Table 9. common structural parameters

KNOT LOCATIONS:	(quoted in units of minute spent over a 10-day sample period) $\{\kappa_{c1} = 0,\ \kappa_{c2} = 28.02,\ \kappa_{c3} = 46.12,\ \kappa_{c4} = 75.33,\ \kappa_{c5} = 109.59,\ \kappa_{c6} = 171.31,\ \kappa_{c7} = 289.31,\ \kappa_{c8} = 1,254\}$						
VARIABLE:	γ_{c1}	γ_{c2}	γ_{c3}	γ_{c4}	γ_{c5}	γ_{c6}	γ_{c7}
Point Est.:	0	9.3340	24.720	147.59	424.56	931.45	1,936.4
90% CI:	—	—	$[24.17, 25.43]$	$[139.5, 159.8]$	$[398.2, 440.9]$	$[876.0, 964.5]$	$[1842.5, 21.88.1]$
VARIABLE:	γ_{c8}	γ_{c9}	γ_{c10}		τ_0	τ_1	φ
Point Est.:	9,969.8	17,626	85,103		6.575	1.058	0.0788
90% CI:	$[9029.8, 10890.0]$	$[14803, 20029]$	$[65550, 97060]$		$[6.468, 6.647]$	$[1.050, 1.065]$	$[0.0744, 0.0820]$

[48] For intrinsic/extrinsic motivation indexes, we included two questions each on the pre-survey and post-survey asking students about their biggest motivations for completing school-related work. Two external motivations were listed alongside two intrinsic motivations, along with a fifth "none of the above" option. We then counted the number of corresponding responses across the four questions and standardize the score by subtracting means and dividing by standard deviations.

Figure 16. Nbhd. SES Impacts on Motivation $\log(\theta_m)$

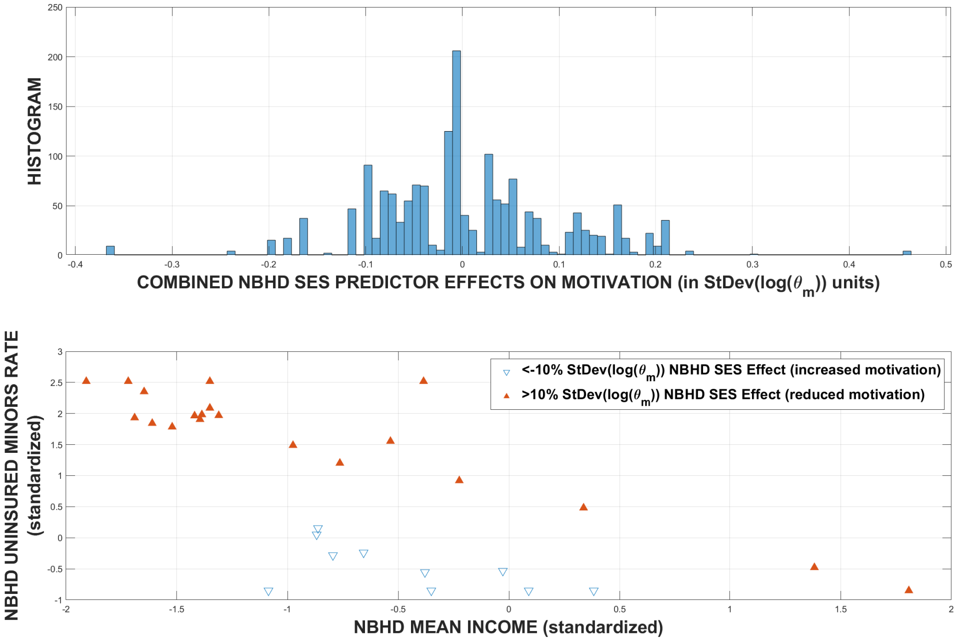
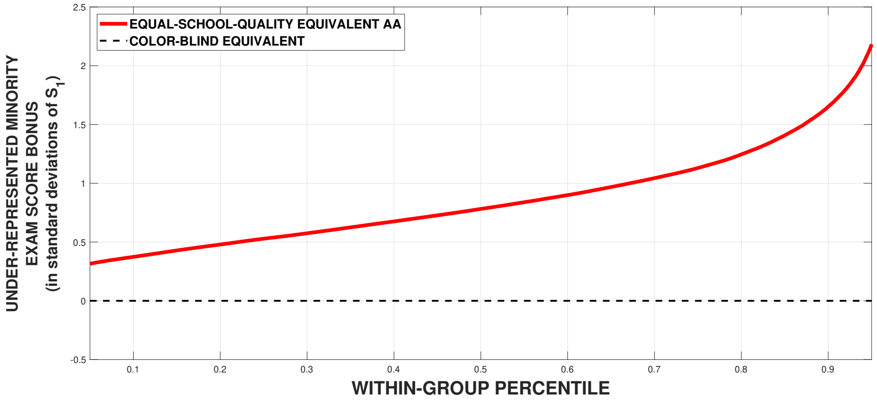


Figure 17. School-Quality-Equalized Affirmative Action



Notes: The figure considers a hypothetical, many-to-many college admissions contest among students in the sample. For $r \in (0.05, 0.95)$ the solid line plots an r^{th}-percentile-specific exam score bonus needed to *exactly offset* handicaps for minority students due to less advantageous school quality assignment relative to their r^{th}-percentile counterparts in the White/Asian group. The dashed line plots the score bonus schedule under a so-called "color-blind" admissions scheme for comparison.

Appendix B. ONLINE SUPPLEMENT: Additional Details

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 5thand 6thgraders. 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 5th/6thgrade 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, and regular homework time play a significant role in predicting total website task completion. These results, and students' various outside options for connectivity (e.g., smartphones or library computers) suggest that internet access is not driving our main empirical results.

B.3. **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

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

Table OS.1. Common Core Focus Areas and 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*.

upper tails of the empirical CDFs of A. The extrapolating B-spline quantile functions overlapped their empirical counterparts to the 85th percentile. We assumed that no student would choose to more than double the available workload on the website, so tails were bounded from above by $A = 160$. We chose a low-dimensional B-spline with 3 knots so that all parameters for the extrapolating quantile functions could be informed by the available data. One advantage of this approach is that we can pre-estimate the

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]
*District*2	-11.16***	*(1.762)*	[-14.62,-7.71]	-8.50***	*(2.168)*	[-12.82,-4.32]
*District*3	-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]
*Contract*2	2.49**	*(1.227)*	[0.08,4.89]	2.44**	*(1.220)*	[0.05,4.83]
*Contract*3	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.

extrapolated upper tails of the work volume distributions, without adding to the computational complexity of the main simulated GMM estimator.

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

[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.

(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 10 and 30 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 1,600 bootstrap samples for the time allocation model, and 500 bootstraps for the Tobit ML model, due to a necessity of estimating multiple specifications of the latter.

For standard errors on student fixed effects, we first bootstrap all common parameters. Then, we combine the bootstrapped parameter samples, $\left\{\tau_0^{(s)}, \tau_1^{(s)}, \varphi^{(s)}, \gamma_c^{(s)}\right\}_{s=1}^{1,600}$, 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.

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 5thand 6thgrades.

B.4. **Supplemental Tables and Figures.**

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

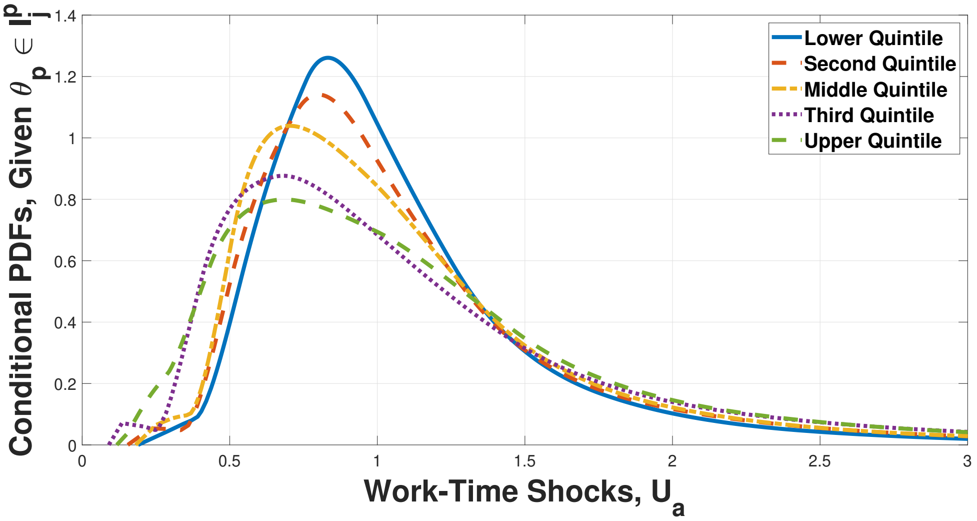


Figure OS.2. CDF Smoothing and Upper Tail Extrapolation

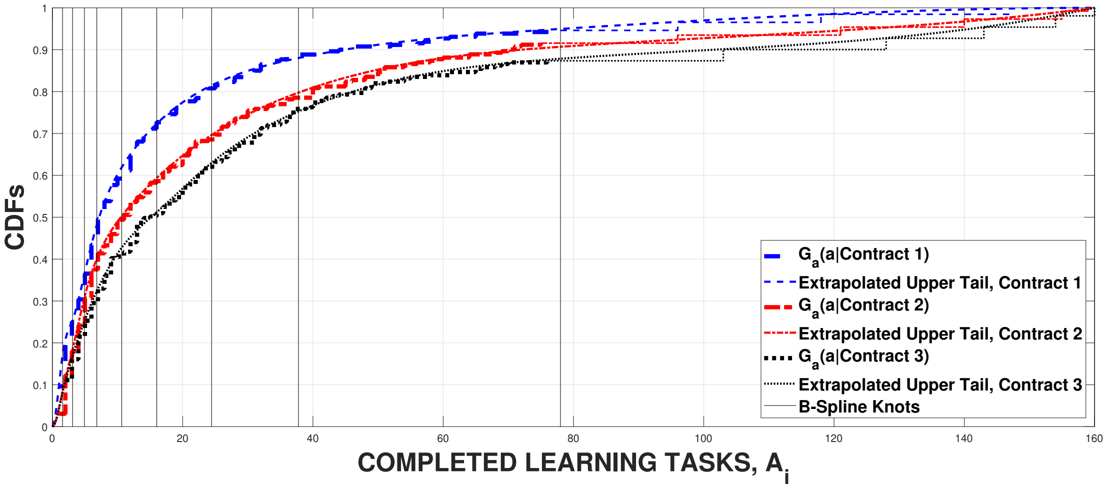


Figure OS.3. Cost Model Fit

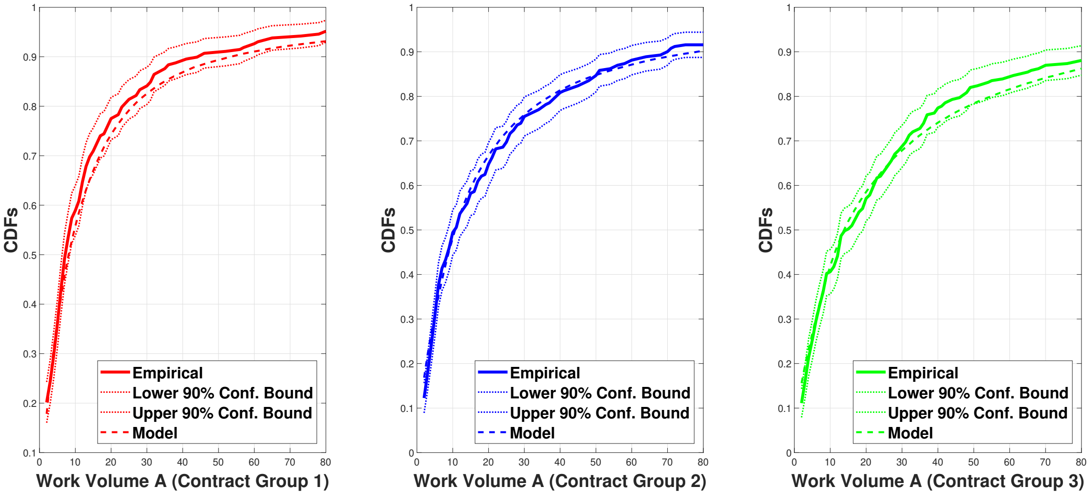
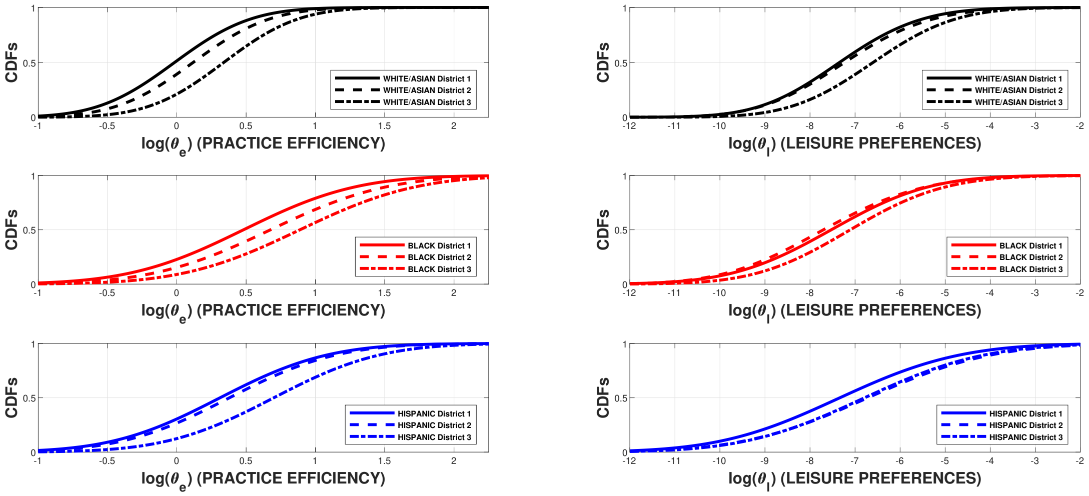


Figure OS.4


CDFs
log(θ_e) (PRACTICE EFFICIENCY)
log(θ_l) (LEISURE PREFERENCES)
WHITE/ASIAN District 1
WHITE/ASIAN District 2
WHITE/ASIAN District 3
BLACK District 1
BLACK District 2
BLACK District 3
HISPANIC District 1
HISPANIC District 2
HISPANIC District 3