NBER WORKING PAPER SERIES

PRODUCTIVITY VERSUS MOTIVATION IN ADOLESCENT HUMAN CAPITAL PRODUCTION: EVIDENCE FROM A STRUCTURALLY-MOTIVATED FIELD EXPERIMENT

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

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

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

This research program would not have been possible without a team of incredibly talented, dedicated, energetic, and tireless research staff. We gratefully acknowledge these men and women (in no particular order) for their pivotal contribution: Andrew "Rusty" Simon, Joseph Seidel, No'am Keesom, Janaya Gripper, Matthew Epps, Justin Holz, Clark Halliday, Debbie Blair, Claire Mackevicius, Allanah Hoefler, Tova Levin, Edie Dobrez, Diana Smith, Kristen Jones, Wendy Pitcock, and an innumerable army of undergraduate research assistants. 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 to extensive feedback from Ariadne Merchant, Daphne Hickman, Morgan Hickman, and Nicholas Merchant as we developed our web-based learning platform. Finally, we also acknowledge particularly helpful conversations with Chris Taber, Samuel Purdy, Mary Moody, Felix Tintelnot, Aloysius Siow, and Joseph L Mullins, as well as feedback from seminar participants at the University of Pennsylvania, the University of Chicago, the University of Wisconsin–Madison, Washington University in St. Louis, Queen's University, and several conferences and workshops. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Productivity Versus Motivation in Adolescent Human Capital Production: Evidence from a Structurally-Motivated Field Experiment

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

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

ABSTRACT

We leverage a field experiment across three distinct school districts to identify key pieces of a structural model of adolescent human capital production. Our focus is inspired by the contemporary psychology of education literature, which expresses learning as a function of the ratio of the time spent on learning to the time needed to learn. By capturing two crucial student-level unobservables—which we denote as academic efficiency (turning inputs into outputs) and time preference (motivation)—our field experiment lends insights into the underpinnings of adolescent skill formation and provides a novel view of how to lessen racial and gender achievement gaps. One general insight is that students who are falling behind their peers, whether correlated to race, gender, or school district, are doing so because of academic efficiency rather than time preference. We view this result, and others found in our data, as fundamental to practitioners, academics, and policymakers interested in designing strategies to provide equal opportunities to students.

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

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

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

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

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

1. Introduction

While the concept of human capital theory can be traced to the writings of Adam Smith, John Stuart Mill, Alfred Marshall and Irving Fisher, until the late 1950's the key factors of production in standard economic models consisted of labor, physical capital, and land (Becker, 1993). Not until Mincer (1958) leveraged human capital to examine inequality in personal incomes did the field of human capital theory begin to take on scientific import.[1] Mincer's work, and subsequent research from the Chicago School and others, unlocked crucial early insights using the human capital approach, including the underpinnings of the growth residual factor, why the ratio of capital to income had decreased over time, and why labor earnings had risen recently despite its stagnation for much of human history (e.g. Schultz, 1961).

This early work set in motion two streams of literature. The first estimates the internal rate of return using variation in human capital based on Becker (1964). The second, based on Ben-Porath (1967), deals with the life-cycle of earnings as individuals trade-off building new human capital versus renting their stock of human capital on the labor market. As these two strands of work make clear, the literature generally considers human capital from the labor market perspective: individuals make investments that develop their skills, and this stock of skills is optimized for generation of income. Empirically, human capital is typically operationalized as being measured in years of schooling completed.

A related line of research on education production functions complements the human capital literature by investigating the *determinants* of human capital (e.g. Cunha & Heckman, 2007; Heckman, 2008; Currie, 2009). In this literature, standardized test scores, or some other survey-based measure of cognitive and/or executive function skills, are interpreted as proxies for skills that are valued on the labor market (see Hanushek, 2020). This body of work can be divided into two distinct periods. The first, following the famous Coleman Report (1966), examined the impact of specific measures of school inputs—e.g., student-teacher ratios, teacher experience, overall school spending, etc.—on student learning (for early contributions see Katzman, 1967; Kiesling, 1968; Bowles, 1970).

The literature has evolved more recently in a second period to focus on an examination of specific aspects of education production, often using data generated via field experiments (e.g. Fryer, Levitt, & List, 2015) or methods concentrating on the effects of teacher quality on test scores (e.g. Chetty, Friedman, & Rockoff, 2014). As Hanushek (2020) points out, this body of research formally links the human capital literature with social science on education production functions. Thus, there is now a useful rationale for interpreting education production estimates as reflecting the long-run economic impacts of educational inputs.

Of course, the field of economics does not have a monopoly on insights concerning skill formation. Foundations of the study of learning can be traced as far back as Plato and Aristotle. According to Schunk (2020), the psychology of learning, influenced by this early philosophical work, began in earnest late in the nineteenth century with James (1890), Dewey (1896), and Titchener (1909) (among others) actively engaged in structuralism and functionalism. The study of human learning expanded during the 20th century, with Bandura (1986, 1997), Bruner (1961, 1966, 1985), Vygotsky (1962, 1978), and others.

In contemporary psychology of education the classical approaches have been replaced by a more sophisticated cognitive model that stresses the influences of a student's perceptions and beliefs on behavior (Carroll, 1962, 1963; Bloom, 1968; Eccles et al., 1983; Wigfield & Eccles, 2000; Eccles & Wigfield, 2002). One particularly influential qualitative framework by Carroll (1962, 1963) began as a modeling exercise

[1] Interestingly, Becker quips that he was quite cautious in using the term "human capital" for the title of his book and thus opted for a long subtitle to avoid criticism (Goldin, 2016).

on learning foreign languages, and highlighted how aptitude, ability, and instruction type interacted to influence a student's choices, and in turn, her level of mastery of a new language (Carroll, 1962). Carroll (1963) extended the model to general learning of any cognitive skill or subject matter. The model postulates five basic classes of variables that account for individual variations in school achievement: aptitude, opportunity, perseverance, instructor quality, and innate ability. Interestingly, while Carroll's qualitative model has been a basis for major programs of scientific innovation in the fields of education (see Denham & Leiberman, 1980) and psychology of learning (Bloom, 1968; Eccles et al., 1983; Carroll, 1989; Wigfield & Eccles, 2000; Eccles & Wigfield, 2002), the economics literature to our knowledge has made no attempt to test or build upon the framework for quantitative research.

A major goal of our study is to draw inspiration from the contemporary learning model in Carroll (1963), as well as its predecessors and successors (e.g. Morrison, 1926; Skinner, 1954; Bruner, 1966; Eccles et al., 1983; Wigfield, 1994; Wigfield & Eccles, 2000), to speak to the human capital and education production function literatures. Our starting point is the emphasis on time as an important variable in skill formation/learning. A focal point of the educational psychology literature is the idea that a child's learning is a function of the time needed to learn and the time actually spent on learning. Under this formulation, Carroll (1963) famously proposed that students accumulate skill by increasing the ratio

$$learning = \frac{time\ spent\ on\ mastering\ a\ concept}{time\ needed\ to\ master\ the\ concept},$$

either by increasing time spent (numerator) or by reducing time needed to learn (denominator), or both. Carroll described the two key parts of the model as "aptitude" (the amount of time a student needs to learn a given task) and "perseverance" (how willing she is to spend time learning the task). Since these two terms have come to hold very different meanings in various social science literatures, we rename the two unobservable student characteristics as "academic efficiency" and "time preference."

We propose a quantitative model of learning that provides direction into the exogenous variation necessary to quantify these two unobservable student characteristics. Our model and experimental design draw upon a novel identification framework proposed by D'Haultfoeuille and Février (2015) and Torgovitsky (2015). Our approach to quantifying academic efficiency relies on standard panel-data methods applied to a remarkably rich dataset on children's time inputs and learning task accomplishment. Following this step, the identification argument for time preferences consists of using exogenous piece-rate incentive variation to derive an empirical mapping between observable student hours worked and their underlying type. This mapping allows us to reverse-engineer a student's cost schedule for supply of would-be leisure time to study, and the distribution of childrens' individual work-time supply costs.

Our research leverages piece-rate incentives since these are the dominant forms of external motivation in academic life: a child is rewarded (or punished) based on how many homework assignments she completes or how many test questions she correctly answers, and not on how much time her homework took or how long she studied for an exam.[2] After structurally estimating the two-dimensional unobserved heterogeneity, we analyze the relationship between the estimated student type parameters and observable factors, including school district, neighborhood characteristics, and demographic variables. This approach allows us to explore, for example, how differences in motivation or study efficiency may contribute to academic performance gaps between different demographic groups. We also can examine how student characteristics differ across the diverse school districts in our sample.

[2]In a sense, this idea is implicit in the Carroll model, though education psychologists focus on learning task accomplishment rather than on piece-rate incentives *per se*.

Finally, we estimate two models of math skill production technology: one focused on short-term learning and one on medium- to long-run learning. This exercise sheds light on how observable factors interact with student type parameters to determine learning gains and overall human capital accumulation. Importantly, this analysis relies on the quantified student unobservables from our structural model in order to solve a classic identification problem of omitted variables and selection bias: do high-performing schools have better outcomes because the school inputs are inherently better, or because more academically adept children tend to self-select onto their rolls? Our estimates facilitate counterfactual analyses of the link between racial achievement gaps and distributions of school quality, which are highly correlated with race in our sample, and in the American population more broadly.

To create the experimental control, data, and variation needed to identify the model and provide relevant policy insights, we must satisfy two necessary conditions: (i) secure a diverse set of school district partners and (ii) design a tool that meticulously tracks student choices and effort under various piece-rate incentive levels. For the first, after months of negotiations we concluded agreements with three diverse school districts in the Chicagoland area that hosted nearly $1,700$ adolescent students in grades 5 and 6 (roughly ages 10 and 11). Importantly, the students come from one high performing/wealthy school district, one middling school district, and one school district that has substantial poverty, lags in operating budget, and where nearly every student metric is well below state averages.

In terms of the second necessary condition, a key feature of our field experiment is that we built a website, accessible only through a login credential assigned to each student, wherein the students could complete up to 80 mathematics modules that we constructed based on professionally developed, age-appropriate math materials. Students had access to the website for 10 days and throughout the process our web server monitored students' activities and tallied successful completion of math modules. Our web-based platform, with its automated, non-invasive tracking and Common Core math materials, was carefully crafted to parallel day-to-day homework activities and a child's associated effort choices. A key to our identification strategy is that we randomized piece-rate incentives for task completion across students, based on the number of modules they completed successfully. Combining this information with pre- and post-experiment measured proficiency using in-classroom mathematics assessments, and a host of other student covariates, we are able to identify the model.

We report several unique insights, which we gather into 3 areas: the student time allocation model, the skill production technology models, and counterfactual analyses. First, our quantitative analog of the Carroll (1963) model contributes 3 novel empirical results to the science of learning. (I) We estimate a remarkably high degree of curvature in a child's utility costs of giving up would-be leisure time for study activity. The key insight is that study-time supply is quite inelastic for all but roughly the 10% most academically inclined students. As a specific policy insight, this result suggests that altering the numerator of Carroll's learning ratio may be a very costly prospect for the average middle-schooler. (II) We observe a remarkable degree of heterogeneity across students in unobserved traits. Monetized utility costs of 3 and 6 hours of foregone leisure time differ by a factor of nearly 7 across the 25^{th} and 75^{th} percentiles among students who were active on our website. In terms of academic efficiency, average time-to-success differed by a factor of 2.7 across the 25^{th} and 75^{th} percentiles among students who were active on our website.[3] (III) Related to the first two results, willingness to forego leisure time is *not* the most important determinant of a student's study effort and learning task accomplishment. Rather, a

[3]Half of our test subjects declined any activity on the website, due to time preference being too high, academic efficiency being too low, or both. Thus, these numbers likely understate the heterogeneity across the overall sample population.

student's academic efficiency, which determines how effectively he can turn inputs into outputs, played a relatively dominant role in shaping academic choices among students in our sample. As a policy insight for an official wishing to increase human capital production, this points to the denominator of Carroll's equation as the location of the proverbial low-lying fruit.

The level of empirical heterogeneity in time preference and academic efficiency also permits an exploration of how these traits relate to observable factors. In the raw data we find racial and gender achievement gaps in standardized math test scores across all 3 school districts. The gaps for race are largely consistent with the literature (e.g. Fryer & Levitt, 2004; Clotfelter, Ladd, & Vigdor, 2009; Hanushek & Rivkin, 2006, 2009; NAEP, 2019) with an interesting data pattern: observed racial gaps are largest in the poorest school district and smallest in the wealthiest school district. The middling school district shows an intermediate gap. Importantly, the race gap in performance is driven by differences in academic efficiency, not time preferences. Indeed, we find either no significant difference in time preferences across racial groups (i.e., between Hispanic and White/Asian students) or differences in time preferences that suggest minority students are *more* motivated than non-minority students (i.e., Black students are more willing to put in time studying than White/Asian students, all else equal).

Considering gender gaps, consonant with the literature (e.g. Hyde et al., 1990; Guiso et al., 2008; Hyde & Mertz, 2009; Fryer & Levitt, 2010; OECD, 2015) we find that males outperform females on standardized math tests, on average, but again there is an interesting trend across school districts. While the gender gap is not evident in the poorest and middling school districts, we find a large gender gap in the wealthiest school district. In terms of its underpinnings, we find that females tend to require less incentive to spend time studying than their male counterparts: their time preference is such that they are willing to spend hours studying, holding external incentives fixed. This difference, by itself, leads to higher academic performance. Yet, in contrast, males tend to have higher academic efficiency, and in net the relative size of this effect compared to observed differences in time preference yields the gender gap observed.

Finally, we find considerable selection on unobservables across the three school districts. Even after controlling for observable student characteristics, there is a 0.76 standard deviation gap in academic efficiency between District 1 and District 3, which is 1.8 times the gap between grade 5 and grade 6 students. Similar patterns are not observed across districts in regards to student time preference. Yet, interestingly, while neighborhood income has little impact on average, we do find that deprivation of non-school developmental resources (e.g., health insurance) is a statistically and economically significant predictor of a child being less motivated for academic pursuits.

A second area of results we report pertains to human capital accumulation and skill formation. We find that both academic efficiency and time preference are important determinants of human capital production, with academic efficiency being roughly three times more important than time preference in determining the initial math proficiency of students. In terms of total factor productivity, we find strong evidence that school quality alters input productivity in an interesting manner: high-performing school districts have higher total factor productivity and lean more heavily on a child's academic efficiency trait, whereas middle- and low-performing schools have lower total factor productivity and lean more heavily on a child's motivation trait in order to generate math skill over time. Furthermore, we find evidence that school quality plays an important role in conversion of learning-by-doing activities into improvements in demonstrated math proficiency.

An important lesson emerges when we pair the first and second area of results. Together, they suggest that educational interventions aiming to decrease gender or racial performance gaps in mathematics by motivating students through incentives or information about the returns to education (such as in Angrist et al., 2009; Fryer, 2011; Fryer et al., 2015; Levitt, List, & Sadoff, 2016; Levitt, List, Neckermann, & Sadoff, 2016; Seo, 2020) might be misguided. This is because such students already tend to be more motivated than their male or White/Asian peers, suggesting that motivation is not the primary barrier limiting their performance. In this spirit, a policy approach based on incentivizing higher effort from such students will struggle to overcome their relative disadvantage.

Our final area of results pertains to counterfactual exercises aimed at investigating the role of access to high-quality education services in explaining racial achievement gaps within our sample population, as well as the potential for incentive-based interventions to ameliorate these gaps. Several interesting insights emerge. First, conditional on key student characteristics, racial differences in school quality account for roughly 45% of the achievement gap between Blacks and Whites/Asians, and roughly 85% of the achievement gap between the Hispanics and Whites/Asians in our sample. Moreover, our model predicts that a leveling of the playing field (in terms of bringing Black/Hispanic school quality up to the same level of White/Asian school assignment) would cause the academically most talented 5% of Black and Hispanic students to actually overtake their 5% most talented White/Asian counterparts in terms of exam score performance.

Second, for policy purposes, we explore two distinct approaches to achieving equality: affirmative action and pecuniary incentives to close the achievement gap. We find that the incentive channel is relatively weak, requiring large amounts of resources to affect outcomes materially. These results strongly suggest that programs or policies to increase the motivation of struggling learners are unlikely to be a cost-effective means of substantially closing demographic performance gaps, since the main driver of these gaps is not a difference in motivation, but rather differences in academic efficiency (driven by factors such as differences in school quality). Moreover, we find that a narrowly-tailored affirmative action scheme that merely un-did the systemically uneven distribution of school quality by race would have to be quite substantial relative to the so-called "color-blind" alternative.

Overall, our results speak to several strands of the literature. First, we clarify and define exactly what is meant by the important unobservable elements, time preference and academic efficiency, in the skill formation context. While the broader literature has used perseverance (e.g. Carroll, 1963), grit (e.g. Duckworth et al., 2007), intrinsic motivation, self-motivation, and other executive function skills (such as in Cunha et al., 2010; Gneezy et al., 2019; Kosse et al., 2020; Cappelen et al., 2020), to describe time preference, our metric is theoretically-driven, clearly defined, and quantifiable. Likewise, while aptitude, cognitive ability, and innate ability have been used to measure academic efficiency, we develop a theoretically-consistent metric that is easily obtained and correlates with key observables. Measuring the two unobserved characteristic types of students is important to both the theorist and policymaker. If the theoretical arguments as to the relative efficiency of different instruments are to be subjected to empirical testing, it is essential to make actual measurements of them. Equally, if education policies are to concern themselves with particular student types or school districts, it is important to understand the optimal approaches and how far a given student can be expected to increase their output by simply increasing time allocation or enhancing academic efficiency. We are unaware of any attempts that have solved this problem, likely because while others have produced careful measurements of some or all

of the inputs and outputs, they failed to combine these measurements into any satisfactory measure of efficiency.

Second, the empirical results sharpen our understanding of a number of crucial concepts in the education literature. For example, we often hear descriptions of unsuccessful students being "unmotivated." Usually in such cases, the criticizer is referring to the fact that the student does not complete homework assignments, show competency on tests, or engage completely within the classroom. Our results and accompanying model call into question the traditional notion of what a "motivated" student is by showing that this logic naively confounds two very different aspects of the child's experience. He may be highly willing to devote an hour (low time preference) to study, or on homework, test prep, and classroom engagement, but if he expects that hour to be unproductive (low academic efficiency) due to lack of high-quality instruction or other resources, then it still may not be rational to exert high effort in response to external incentives because the time needed to perform well is unreasonably high. Or, he might be very motivated, putting forth high effort, but because of low academic efficiency he appears unmotivated.

In addition, our view that academic efficiency is the amount of time required by the student to develop skills means that given enough time all students can conceivably obtain key skills. Under this reasoning, learning is available to all, we just need to find the means to help each student. Our particular formulation has fundamental implications for education, and guides us to recast the education problem from one of a goal of equal achievement for all to one of equal opportunity for all, ensuring that anyone who is willing to put in the time and work hard has the potential to succeed. Under this view, we need to understand what policy approaches provide such student equality, rather than focus primarily on equity considerations. A related literature in this spirit is the literature on school district quality and moving to opportunity. Much of this literature focuses on interventions that result in children relocating to new schools at some time during their studies, and is therefore often focused on the disruptive effects of moving or changing school environments (e.g. Katz, Kling, & Liebman, 2001; Hanushek, Kain, & Rivkin, 2004; Rumberger, 2015; Chetty, Hendren, & Katz, 2016; Schwartz, Stiefel, & Cordes, 2017; Cordes, Schwartz, & Stiefel, 2019; Schwartz, Horn, Ellen, & Cordes, 2020).

In an immediate policy sense, our findings have implications for the design of programs to close achievement gaps across demographic groups. By pinpointing the underpinnings of skill deficiencies, we learn that many students who appear unmotivated and do not complete assignments are likely no less willing to put in time studying than their more-successful peers, but that academic success (or even progress) is difficult for them to achieve. This low academic inefficiency in turn discourages them from investing time in their studies. When we consider performance gaps across demographic groups more broadly, we show how these gaps are not due to differences in motivation—in fact, the motivation gap either plays no role or even points in the opposite direction—but rather, are due to differences in how efficiently students in different demographic groups convert study time into the successful completion of academic assignments and learning gains.

This insight highlights that under-represented minority students are struggling compared to their peers, not because they are unwilling to spend time studying, but because they are more likely to lack the foundation (e.g., literacy and numeracy skills, study skills, high-quality school inputs, support at home) on which academic progress can be built more easily. This means that initiatives to close performance gaps by increasing motivation among under-performing groups, whether through information or incentives, are very much not addressing the primary barriers holding these groups back. Such programs

that encourage greater effort from marginalized groups, who we show are already at least as willing to put in time studying than others, is unlikely to substantially close any performance gaps. Real change will more likely come from efforts to better understand and address the reasons why these groups are less able to effectively convert their study efforts into learning gains. We show that much of the racial performance gap in mathematics comes from differences in school quality and resource deprivation due to poverty, which may influence their foundational literacy and numeracy skills, limiting learning and discouraging effort, even among those eager to learn.

The remainder of our study is structured as follows. Section 2 outlines the quantitative theoretical framework that underpins our research design and empirical strategy. Section 3 discusses model identification, with an emphasis on how experimental variation can enable us to generate the requisite set of observables to uniquely quantify unobserved student characteristics and other structural primitives. The identification argument provides a number of insights regarding how the experiment must be designed to achieve the proper variation in the data. Section 3 also presents our research design, focusing on the crafting of an organic learning scenario, incentives variation, and how subjects were chosen. Section 4 presents estimator details. Section 5 presents both reduced-form and structural results. Section 6 presents a counterfactual simulation exercise to explore variation in school quality as well as two distinct public policies that have been used to lessen racial and gender achievement gaps: affirmative action and pecuniary incentives. Section 7 concludes. An appendix contains additional technical details, graphs, and tables.

2. Theoretical Framework

Causal inference on educational outcomes has always been impeded by the canonical identification dilemma of unobserved student characteristics. It is well documented that students who attend better-resourced primary, secondary, and post-secondary schools have better academic outcomes such as grades, exam scores, college placement, jobs, etc. What is much less understood is the extent to which these better outcomes are driven by selection of students who would have been high achievers *anywhere*, or whether differences in actual school quality are responsible for observed achievement gaps. In the United States, with K-12 education financed primarily by local property taxes, this confounding of selection and treatment effects is particularly daunting from a researcher's perspective, and yet particularly important to understand for policymakers. We take a novel approach to solving this problem by developing an empirical framework that allows the researcher to individually quantify students' unobservable characteristics related to both motivation and underlying learning ability. This advance, in turn, allows for these characteristics to be included as explicit controls when investigating the role of school quality in shaping student proficiency.

Our quantitative model of knowledge and skill formation is closely related to the qualitative expectancy-value theory used extensively to assess study effort in the education and psychology literatures (e.g., Carroll, 1963; Eccles et al., 1983; Wigfield, 1994; Eccles & Wigfield, 2002; Wang & Degol, 2013). The two key components of the expectancy-value model of achievement are (i) a student's perceptions about her ability at a particular task, and (ii) a combination of the intrinsic value and cost they experience while engaging in the task (Wigfield & Eccles, 2000). This literature typically assesses these characteristics using Likert-scale survey questions about how good a student is at learning new concepts in math and how much they enjoy working on math relative to other activities. Here we propose a framework for estimating these characteristics from observable behavior using the principle of revealed preference. The

two primary parameters in our model are similar to the expectancy-value model: students' achievement in math is a function of their ability to complete learning tasks and their perception of the costs/benefits of allocating time to math relative to other activities. Our contribution is in formalizing a quantitative model of student choice and proposing a new method to elicit these parameters in a way that is more closely tied to individual decisions, thus enabling informative counterfactual analysis.

2.1. **Unobserved Student Heterogeneity.** Our model formalizes adolescent skill formation as the result of a production process whose form may vary by school attended or other environmental factors. Individual student characteristics serve as the principal productive inputs in this process. For simplicity of discussion we focus on subject-specific proficiency in mathematics, though it is straightforward to generalize beyond a single subject. Each student is characterized by two unobserved traits: *academic efficiency*, denoted θ_e—which governs the idiosyncratic rate at which learning-by-doing tasks are accomplished—and *time preference*, denoted θ_l—which governs a child's motivation or willingness to substitute a fixed unit of time away from the next best option and toward math activities. Both characteristics represent costs in either the time or utility dimension, so that higher values of θ_e imply more time required to complete a given learning task, and higher values of θ_l imply greater dis-utility of spending time practicing math.

When formalizing the roles of these two characteristics it is important to recognize that piece-rate incentives are the predominant mode of reward and punishment in real-world educational settings. For example, students are rewarded with good grades or exam scores based on how many homework assignments they complete, or how many questions they answer correctly in a timed examination. Conditional on homework completion or exam performance, these rewards are unaffected by how many hours of homework or study time it required. As such, academic labor-leisure choices are not dictated solely by time preferences: holding θ_l and external piece-rate incentives fixed, a reduction in θ_e implies that each unit of a student's time is more valuable. Therefore, both student traits play a central role in decision making.

Idiosyncratic differences in θ_l 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 on mathematics problems. Heterogeneity in θ_e may reflect either differences in a child's initial proficiency level, or differences in a child's study process, academic support network, or innate ability that affect how quickly she regularly completes assignments. Since both traits are a mixture of innate and environmental components, for each student i we allow them to evolve with changing circumstances according to the following

$$\log(\theta_{ei}) = \boldsymbol{X}_{ei}\boldsymbol{\beta}_e + \eta_{ei}, \quad \text{and} \quad \log(\theta_{li}) = \boldsymbol{X}_{li}\boldsymbol{\beta}_l + \eta_{li}, \tag{1}$$

where $\boldsymbol{X}_{ei} = [1, x_{e1i}, \ldots, x_{ek_ei}]$ and $\boldsymbol{X}_{li} = [1, x_{l1i}, \ldots, x_{lk_li}]$ are vectors of environmental characteristics including school quality, family academic support, socioeconomic variables and other factors. The η_{ei} and η_{li} terms represent the truly idiosyncratic portions of student i's unobserved traits $(\theta_{ei}, \theta_{li})$. Going forward, we assume the following about the joint distribution of unobserved student traits:

Assumption 1. *The two idiosyncratic components follow a bivariate normal distribution* $(\eta_{ei}, \eta_{li}) \sim \text{BVN}(\mathbf{0}, \boldsymbol{\Sigma}_i)$*, where the variance-covariance matrix* $\boldsymbol{\Sigma}_i$ *may potentially vary by observable student characteristics.*

It is important to understand that the notion of human capital itself has several distinct aspects. While $(\theta_{ei}, \theta_{li})$ may be considered forms of human capital themselves, they represent a collection of factors outside of the student's control, at least over a short-run horizon. However, $(\theta_{ei}, \theta_{li})$ govern decisions

and rate of progress in the short-run which are under a child's control, and over time accumulate into her stock of invested human capital. Forms of invested human capital are often measured by or reflected in demonstrated ability on standardized assessments. Measured mathematics proficiency, θ_{ei}, and θ_{li} are all different aspects of human capital: the first reflects one's current stock of task-specific skill and the latter two govern one's ability to acquire new task-specific skill. Moving forward, we describe this distinction by using the terms *math skill* or *proficiency* to refer to demonstrated performance on standardized assessments, and the term *characteristics* to refer to a student's underlying type variables, $(\theta_{ei}, \theta_{li})$. The maintained assumption behind our framework and research design is that characteristics can be treated as fixed over the short-run, while skills and proficiency are malleable over relatively shorter periods of time.

2.2. **Student Choice Model.** Consider child i's choice of study time and the resulting volume of learning tasks that are successfully completed with that time investment. These are endogenously determined by a decision process that hinges on both θ_l *and* θ_e in the presence of piece-rate incentives. Formally, achieving gains in math proficiency over the short-run is a process of performing repeated, discrete, learning-by-doing tasks (e.g., homework assignments). Each student chooses Q_i, representing how many learning activities to complete. Since Q_i is a means to an end (i.e., expanding one's permanent skill set in mathematics), we will sometimes refer to it as *interim output*. A piece-rate payoff function $P_i(q)$ represents the external benefits received by student i from completing q units of learning activities. Of course, producing interim output requires that the student give up some quantity of her time T_i, which could otherwise be used for the best outside option (e.g., video games, sports, other tasks, socializing with friends, etc.). Academic efficiency θ_{ei} shapes the mapping between T_i and Q_i through the following relation: $T_i(Q_i) = \sum_{q_i=1}^{Q_i} \tau_i(q_i; \theta_{ei})$, where

$$\tau_i(q_i; \theta_{ei}) = \theta_{ei} \times \tau_0 \times q_i^{-\gamma} \times u_{q_i}, \qquad \theta_{ei}, \tau_0, \gamma, u_{q_i} > 0. \tag{2}$$

In equation (2), $\tau_i(q_i; \theta_{ei})$ represents the time spent by student i on completing her q_i^{th} unit of interim output. $\tau_i(\cdot)$ has several components: τ_0 is the mean initial production time on the first unit across all individuals, while the term $q_i^{-\gamma}$ is an *experience curve* that allows for a student's rate of progress to increase with additional work (when $\gamma > 0$) or for it to deteriorate through exhaustion (when $\gamma < 0$). The student's academic efficiency θ_{ei} scales this mean production curve up or down, relative to her average classmate, and u_{q_i} is a transitory *iid* shock to production time, representing unpredictable fluctuations in difficulty level across tasks, mental state, distractions, etc.

Assumption 2. *The (potentially heteroskedastic) unit-specific production shock U_{q_i} follows distribution $F_{u|\theta_{ei}}(u|\theta_{ei})$ with continuous density $f_{u|\theta_{ei}}$ that is bounded away from zero on support $[\underline{u}, \overline{u}] \subset \mathbb{R}_+$.*

Finally, student i experiences dis-utility of shifting time from the outside option to math study according to the following differentiable cost function

$$C(T_i; \theta_{li}) = \theta_{li} c(T_i), \quad \theta_{li} > 0, \tag{3}$$

where dis-utility is denominated in the same units as the piece-rate payoff schedule $P_i(q)$. Note that multiplicative separability is a non-trivial assumption in the model, as it will be central to our identification strategy (see Section 3 below). We also assume the following regularity conditions to ensure a well-behaved decision problem for student i:

Assumption 3. *Costs and marginal costs are increasing,* $c'(t) > 0$*, 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* $c'(0) = 1$*.*

Combining the costs and benefits of practice activities implies an optimal stopping problem. After successfully completing $(q-1)$ learning tasks, a student decides the maximum time t^* she is willing to spend on the q^{th} unit of interim output, according to

$$t^* \equiv \underset{t \geq 0}{\operatorname{argmax}} \Big\{ Pr\big(q^{\text{th}}\ success | t, (q-1)\big) \big[P_i(q) - P_i(q-1)\big] - \big[\theta_{li} c(T_i(q-1) + t) - \theta_{li} c(T_i(q))\big] \Big\}. \quad (4)$$

In words, after completing each unit of interim output, student i makes a mental calculation of how much additional work would cause the marginal cost of additional time to swamp the marginal benefit of one more successfully completed learning task. Here, the probability that she succeeds on task q given t units of time spent depends on the distribution of the production shock U_{q_i}: $\Pr\big(q^{\text{th}}\ success | t, (q-1)\big) = F_{u|\theta_{ei}}\left(\frac{t}{\theta_{ei}\tau_0(q-1)^{-\gamma}} \middle| \theta_{ei}\right)$. If she is able to achieve the q^{th} success with some work time $t < t^*$, then she re-optimizes decision rule (4) with a comparison of q versus $(q+1)$ achieved successes, and continues on. Otherwise, she discontinues effort and the final values of T_i and Q_i that enter her short-run incremental production function are determined by her optimal stopping point.

The above model makes it clear why both student traits θ_e and θ_l contribute to a students' supply of her time to math studies. Academic efficiency θ_e determines how burdensome a given level of achievement is in the time dimension, and θ_l determines how costly the expended time and effort are in the utility dimension. In public debate about academic policy, students are often labeled as "more motivated" when they complete more homework assignments on time, but the model illustrates how this way of thinking actually conflates two very different aspects of the student experience when piece-rate incentives are in play. Student i may be highly motivated relative to student j in the sense of willingness to re-allocate leisure time toward math activities (i.e., $\theta_{li} < \theta_{lj}$), and yet may still complete fewer homework assignments if the academic efficiency difference between them ($\theta_{lj} - \theta_{li} > 0$) is large enough, due to asymmetric resource allocations, such as school quality, tutors, support network, etc.

From a policy perspective, when we observe a student performing poorly on an exam, this may be due to either high time costs (i.e., high θ_l), a lack of foundational math and study skills (i.e., high θ_e), or some combination of the two. A deeper understanding of how these two factors interact at the student level may help practitioners to achieve more efficient, individually-tailored allocation of scarce resources: do Bobby or Suzie need tutors, or do they simply need someone to convince them that math is enjoyable, relevant, or at least not onerous? At the group level, understanding the distributions of these two characteristics and their relation to educational resources can produce crucial insights for policymakers interested in remediation of demographic achievement gaps.

2.3. **Initial Math Skill.** Since (θ_e, θ_l) determine a child's short-run choices and task accomplishment which accumulate into long-run outcomes, we model a student's initial math proficiency level S_i as the outcome of a Cobb-Douglass production process with θ_{ei} and θ_{li} as its principal inputs. Specifically,

$$S_i = A_i \times \theta_{ei}^{\alpha_{ei}} \times \theta_{li}^{\alpha_{li}} \times \epsilon_i, \quad (5)$$

where A_i is total factor productivity (TFP), and ϵ_i is an idiosyncratic, multiplicative shock. Total factor productivity and the Cobb-Douglas production shares $(\alpha_{ei}, \alpha_{li})$ are allowed to be idiosyncratic, depending on observable student covariates:

$$\log(A_i) = \boldsymbol{W}_i \boldsymbol{\alpha}_0, \quad \alpha_{ei} = \boldsymbol{W}_i \boldsymbol{\alpha}_e, \quad \text{and} \quad \alpha_{li} = \boldsymbol{W}_i \boldsymbol{\alpha}_l, \quad (6)$$

with $\boldsymbol{W}_i = [1, w_{1i}, \ldots, w_{ki}]$ including school quality, family learning support, socioeconomic variables, and other factors.[4] The error term ϵ_i accounts for the cumulative impact of transitory shocks to the production process over time as well as noise in the exam instrument used to measure a student's math proficiency level.

2.4. **Incremental Gains in Math Skill.** Over a short-run horizon—a period spanning weeks—we propose a separate but related production model in which student i's study time T_i and successful completion of learning tasks Q_i contribute to gains in her mathematics proficiency level. Let ΔS_i denote the short-run improvement in a student's measured math proficiency,

$$\Delta S_i = \Delta_{0i} + \Delta_{1i} T_i + \Delta_{2i} T_i^2 + \Delta_{3i} Q_i + \Delta_{4i} Q_i^2 + \Delta_{5i}(T_i \times Q_i) + \varepsilon_i, \tag{7}$$

where ε_i is an idiosyncratic, transitory shock. Similarly as in initial math skill production, the short-run production parameters depend on a vector of individual covariates,

$$\Delta_{ji} = \boldsymbol{V}_i \boldsymbol{\delta}_j, \quad j = 0, \ldots, 5, \tag{8}$$

where $\boldsymbol{V}_i = [\boldsymbol{W}_i, S_i, \theta_{ei}, \theta_{li}]$. Note that by including unobserved student traits in $\boldsymbol{V}_i$ we are allowing for them to play a dual role in shaping a child's ability to acquire new skill: first, they underlay choices of T_i and Q_i, and second, they may alter the rate at which learning activities translate into knowledge gains. Including S_i as a control allows for the possibility of a decreasing-returns-to-scale technology where incremental gains of a fixed size become more difficult as a student achieves greater subject mastery.

Note that our student choice model provides a micro-foundation for our model of short-run skill formation, where $(\theta_{ei}, \theta_{li})$ determine (T_i, Q_i), which in turn drive incremental gains in i's measured mathematics proficiency. The short-run skill formation technology is also consistent with the long-run technology for initial math skill: both are fundamentally driven by the interplay between individual student inputs and developmental resources of various types. The biggest difference between model (5) and model (7) is that fine-grained information on time inputs and task accomplishment are feasible for researchers to collect over short-run horizons, but much more difficult over longer spans of time. Therefore, in absence of ideal observables, model (5) uses student traits $(\theta_{ei}, \theta_{li})$ as a stand-in for the terms $(\theta_{ei}, \theta_{li}, T_i, Q_i)$ in model (7). Of principal interest for policymakers is the question of school quality, which may impact student outcomes through three distinct channels in our model: (i) it may influence the long-run evolution of student characteristics θ_{ei} and θ_{li}, (ii) school quality may have a direct impact on the level of math skill development (through the intercept terms in equations (5) and (7)), and (iii) it may indirectly alter the manner in which the production technology converts its primary inputs into new learning (through the slope terms in equations (5) and (7)).

3. Research Design

3.1. **Experimental Motivation and Causal Identification Overview.** Our research design builds on our study-choice and skill-formation framework to bring together experimental and structural methods to quantify unobserved student characteristics. Our strategy uses the student choice model as a basis for an econometric framework, where field experimental methods shape a data-generating process with the requisite sets of observables and variation to enable identification of the structural parameters (θ_e, θ_l) at the individual level. This data-generating process is also carefully crafted to be as true to students'

[4]Substituting equation (6) into equation (5), the long-run production model is equivalent to a regression of $\log(S_i)$ on θ_{ei}, θ_{li}, $\boldsymbol{W}_i$, and a complete set of pair-wise interactions between $(\theta_{ei}, \theta_{li})$ and the variables in $\boldsymbol{W}_i$.

everyday academic choices and experiences as possible. With this in mind, we conducted the field experiment among 5th- and 6th-grade students in three Illinois school districts. We offered varying monetary incentives for completion of extra-curricular learning activities on a math study website that we developed. Our approach to quantifying unobserved student traits builds on standard panel-data methods (for θ_e), and on recent advances by Torgovitsky (2015) and D'Haultfoeuille and Février (2015) on the use of discrete instruments to identify continuous, individual-level heterogeneity (for θ_l).

To develop basic intuition for how our method quantifies two-dimensional student traits, consider a hypothetical "ideal" experiment (from a research perspective) where feasibility constraints are non-existent. Consider two students, *Bobby* and *Suzie*, who perform poorly on a standardized math exam. The exam score alone indicates that each student is struggling, but it does not offer insights as to why. To answer this question, the researcher first obtains identical copies of the two students, call them *Bobby** and *Suzie**—i.e., identical in biology, ability, preferences, attitudes, etc.—and 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 mathematics textbook, and there is also a couch with a TV and a video gaming system, a smart phone connected to social media, and other leisure opportunities. Upon entering the observation room, the researcher presents piece-rate wage offer p to *Bobby* and *Suzie* and $p^* > p$ to *Bobby** and *Suzie** for working through a series of discrete math assignments and demonstrating proficiency in each according to some well-defined 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 further that over 2 weeks *Bobby* and *Suzie* complete 5 and 10 math assignments, respectively, whereas *Bobby** and *Suzie** complete 7 and 13. The research team measures average rates of progress across math assignments for each child, and can infer $\theta_{e,Bobby}$ and $\theta_{e,Suzie}$ as student fixed effects. This information implies effective mean hourly wage rates for each of the four children. For example, suppose that, given *Bobby*'s average rate of progress, his effective hourly wage rate is \$15/hour, whereas *Suzie* works somewhat slower and has an effective hourly wage rate of \$12/hour instead. Note that all differences in mean hourly wage between *Suzie* and *Suzie** are due solely to their piece-rate offers $p < p^*$, since they are identical and have the same $\theta_{e,Suzie}$ trait. Since *Suzie*/*Suzie** produced more output than their same-piece-rate counterparts *Bobby*/*Bobby**, this is an indication that *Suzie* is more easily motivated to allocate time from other activities to math than *Bobby* (i.e., $\theta_{l,Suzie} < \theta_{l,Bobby}$).

More concretely, the hourly wage differences under p and p^* can be used to compute labor-supply elasticities for *Bobby* and *Suzie*, respectively. With this information in hand, and since $\theta_{l,Bobby}$ and $\theta_{l,Suzie}$ 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 its form, independent of *Bobby*'s and *Suzie*'s idiosyncratic traits. For example, *Bobby*'s output increased by 40% while *Suzie*'s output under the same proportional wage increase rose by only 30%, indicating that marginal costs must be higher from *Suzie*'s baseline output of 10 assignments, relative to *Bobby*'s baseline of 5 assignments. Moreover, feasible inference on the form of the common cost schedule become richer as the experiment is repeated with an increasingly larger set of *Bobby*'s and *Suzie*'s classmates, *Jill*, *Tommy*, etc. With a complete picture of the shape of the common cost schedule $c(t)$, the researcher can then separately infer each child's individual leisure preference $\{\theta_{l,Bobby}, \theta_{l,Suzie}, \theta_{l,Jill}, \theta_{l,Tommy}, \ldots\}$.

While informative as a thought exercise, much is obviously infeasible or unethical about the above "ideal" experiment. However, using field experimental methods and modern web-based technologies, one can capture the essential elements of the above scenario while maintaining a level of realism and familiarity that would be impossible within a controlled laboratory setting. Rather than cloning students, one can easily clone groups of students through individual-level randomization. This ensures that, while no two groups will contain identical copies of the same child, the overall distributions of unobserved characteristics will be the same.

Similarly, rather than sealing students into observation rooms, one can move extra-curricular learning materials online, where a web server can meticulously monitor activities in a much less invasive way. One challenge to this web-based alternative is that the researcher cannot control for the role of a student's regular educational activities such as classroom instruction and graded homework assignments for school. However, this does not threaten model identification *per se*, provided that the distributions of regular educational activities are uncorrelated with treatment assignment.[5] Rather, it merely changes the interpretation of the structural parameters somewhat. In the hypothetical, infeasible experiment above, a child's willingness to allocate time toward math activity is judged relative to the baseline of *zero activity*, while in our web-based setup θ_{li} represents marginal willingness to allocate *extra time on the margin*, above and beyond their regular schoolwork.

The web-based tracking setup has two considerable advantages as well. It allows students to engage in academic decision-making against the backdrop of the myriad outside options for their time—sports, clubs, music activities, informal play with friends, chores, etc.—that form a natural part of their normal life routine. Our web-based research design also provides a general proof of concept for powerful new diagnostic tools cheaply available to education practitioners at scale given recent, dramatic increases in K-12 educational materials being moved to online formats.

In the following sections, we provide specific detail on the design of our field experiment, including recruitment, randomization, incentive variation, math proficiency assessment, website structure, and data collection. Our intuitive discussion above also glosses over an important issue of sample selection: how would identification be affected if *Bobby* spent no time on math under p, while his alter-ego *Bobby** produced positive interim output under p^*? Holding piece-rate incentives fixed, there will be a region of 2-dimensional characteristic space where either θ_l or θ_e (or both) are prohibitively large to rationalize any amount of positive effort. To solve this problem, we use Assumption 1 on joint log-normality, further discussed in Section 4 below, to perform a sample-selection correction which extrapolates into the unidentified region similarly as the traditional method pioneered by Heckman (1979).

3.2. **Study Sample.** We partnered with three public school districts in the Chicago-Naperville-Elgin MSA during academic year 2013-2014. A total of 1,676 5$^{\text{th}}$- and 6$^{\text{th}}$-grade students participated in the experiment, with 46% of them drawn from *District 1*, and 27% each coming from *District 2* and *District 3*.[6] The three districts differed widely by local population and administrative characteristics. These differences are described in Table 1. Relative to the state of Illinois, which is demographically most representative of the national population among all 50 U.S. states, District 1 was above-average on faculty

[5]Individual randomization ensures that treatments are independent of school. A possible threat to identification would be if students responded to extra-curricular incentives by neglecting regular schoolwork. We could not access children's academic records due to privacy concerns, but in multiple conversations administrators and teachers universally expressed a strong impression that no reduction in homework completion rates occurred during the sample period. We also find evidence in our survey data consistent with their reports (see Section 3.4 below).

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

TABLE 1. 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 rounded to the nearest 10 pp and represents the average percentage across $5^{th}$and 6^{th}grades.

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 its overall financial resources per pupil. District 2 was remarkably close to the state averages on these dimensions, while District 3 lagged considerably in terms of student academic performance. This was despite District 3 having higher than average per-student operating budget, but this budget also includes spending on social workers, guidance counselors, building maintenance, lunch subsidies, non-instructional support programs, etc.

The populations these three districts serve are similarly ordered in terms of socioeconomics. District 1's student population is substantially more affluent by both income and wealth—with all but 15% of its operating budget derived from local property taxes—and has only a negligible burden of teaching curriculum to children with limited English language proficiency. District 2 is once again closest to the state means, while District 3 is considerably less affluent by income and wealth, and has a relatively large fraction of students with limited English language proficiency (including many Hispanic immigrant families). Finally, the other striking difference across the districts is the racial profiles of the communities they serve (see Table 2 below). District 2 has a racially diverse student body, while District 1 has few Black or Hispanic students and District 3 is almost entirely comprised of Blacks and Hispanics.

3.3. **Field Experiment Details.** We worked closely with 5^{th}- and 6^{th}-grade math teachers across the three participating school districts to implement the field experiment. The major research advantage to this partnership was that participation in the study was on an opt-out basis, allowing the research team to achieve a sample that was much more representative of the local populations our partner schools serve.[7]

[7]Prior to the study, parents were informed and given the opportunity to opt their child out of participation. On the first day of the study, when a diagnostic math pre-test was given in class, individual students were also given opportunity to opt

A primary feature of the experiment was a website on which the students could complete up to 80 mathematics modules, referred to as "quizzes," across five general topics. Students had access to the website for 10 days and could complete as many of the quizzes as they chose. Throughout the process, our web server monitored students' activities and tallied successful completion of quizzes. Piece-rate incentives were offered for task completion on the website, based on the number of quizzes completed successfully, rather than on time spent. We also measured proficiency using in-classroom mathematics assessments. This section provides more specific details about the experimental process.

3.3.1. *Math Proficiency Assessment and Other Student-Level Data.* Prior to being randomly assigned into a treatment group, 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 6$^{\text{th}}$-graders.[8] 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 5$^{\text{th}}$- and 6$^{\text{th}}$-graders to the same materials. This facilitated an even comparison between age groups, allowing us to cleanly estimate the effect of an additional year of schooling on skill formation.

We used Common Core Math Standards definitions to categorize each problem into one of 5 subject categories: (i) *equations and algebraic thinking*, (ii) *fractions, proportions, and ratios*, (iii) *geometry*, (iv) *measurement and probability*, and (v) *number system*.[9] For the pre-test and post-test, we randomly selected a large subset of problems from the math question bank and further categorized them as *easy*, *medium*, or *hard*, depending on their complexity level or number of steps required to solve. Finally, to ensure uniformity of subject content and difficulty level, both the pre-test and post-test were populated with similar sets of 36 questions: 8 each from subjects (i), (iii), and (v), and 6 each from subjects (ii) and (iv). Of the 36 questions, 20 were selected from 6$^{\text{th}}$-grade materials and the other 16 from 5$^{\text{th}}$-grade materials, and the easy, medium, and hard categories were represented by 15, 12, and 9 questions respectively, spread evenly across each exam. We computed pre-test scores S_{1i} 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

themselves out of participation. Parents and students appeared generally enthusiastic about the study, and opt-out rates were negligible ($< 5\%$) across all schools and classrooms partnering in the study.

[8]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).

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

home, fraction of peer social time under adult supervision, and enrollment in organized sports, music activities, and/or clubs). We also gathered socioeconomic indicators from the American Community Survey for each of the ≈160 (rounded to nearest 10 to preserve anonymity) US Census block groups where our test subjects resided, each of which can be thought of as a neighborhood. Within each neighborhood we collected mean household income (a proxy for affluence), and the fraction of minors with no private health insurance (a proxy for deprivation of non-school developmental resources).[10]

3.3.2. *Website Structure.* Our website was accessible through a login credential assigned to each student.[11] This meant that our web server could automatically track activities and measure progress for each child without affecting user experience in any perceivable way. The primary component of the website was a set of 80 math modules, each consisting of a set of 6 multiple-choice questions from our bank of age-apropriate materials.[12] The passing criterion for successful completion of each quiz was at least 5 out of 6 questions answered correctly. Each student was allowed unlimited attempts at each quiz, but for each new attempt the ordering of the questions and the ordering of the choices was randomly perturbed. Adolescent pilot study participants reported a feeling 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 modules on the website were organized into 55 general-topic quizzes (with balanced portfolios of the 5 math topics mentioned above), and 25 topic-specific quizzes (5 per topic). Aside from balancing topical content, math questions were selected at random from our bank of math problems, so that relative difficulty was impossible to predict from one quiz to the next. After each quiz attempt, an automated, interactive feature provided optional feedback, which the student could choose to skip through or learn from.[13] The web server tracked time spent on each quiz (across all attempts) by recording a timestamp for each unique page view. Since only one math problem appears per page view within each quiz, this resulted in a high-frequency log of work times for each child.[14] The website logged successful completions into a database, and visually tracked current earnings and progress for the user by color-coding passed quizzes differently from those not passed. The site also included a prominent reminder of the child's piece-rate incentives. Through a combination of these capabilities

[10]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.

[11]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 tech support email throughout the study, with someone on-call 24/7 to quickly resolve any login problems. These turned out to be few, given the intuitive nature of the login credentials.

[12]Six was chosen because adolescent pilot study participants generally expressed the feeling that more than 6 was too much for the piece-rates we had in mind.

[13]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.

[14]One technical concern was how to deal with a small number of spurious page-view times that resulted when a child closed her web browser in the middle of a quiz attempt without logging off. We truncate this small number of spurious work time observations using a simple adjustment proposed by (Cotton, Hickman, & Price, 2020a, Online Appendix) based on failures of a full support condition in the subject-specific work-time distributions.

and students' labor-leisure choices, we were able to derive our principal observables: Q_i, total quizzes passed by student i; T_i, total time worked; and $\left\{\{\tau_{q_i}\}_{q_i=1}^{Q_i}\right\}_{i=1}^{N}$, a panel of student-unit work times.

Some important distinctions between our website data and our in-class mathematics assessment data are worth emphasizing. Although information collected from both sources measures some aspect of a child's rate of task progress through time, exam scores reflect proficiency in a controlled (i.e., subject to time constraints) and un-monitored (i.e., with no real-time feedback) environment whereas website data reflect rate of progress through quality-monitored practice activities in absence of binding time constraints. Thus, the measures derived from these two data sources—current skill stock versus academic efficiency—represent distinct aspects of a child's learning process.

3.3.3. *Piece-Rate Incentives and Randomization.* Our experimental design centers on randomized incentives. We adopted a linear piece-rate schedule $P_i(q) = (b_i + p_i q)\mathbb{1}(q \geq 2)$ with a constant marginal piece-rate that would be easy for adolescents to understand. We varied both the base payment b_i, for showing up and completing the minimum amount of work, and the marginal piece-rate payment p_i. No payments are offered until a child has passed 2 quizzes, which ensures a within-student panel for each individual i. Each student was randomly assigned to one of three possible contract groups: $(b_1^*, p_1^*) = (\$15, \$0.75)$, $(b_2^*, p_2^*) = (\$10, \$1.00)$, and $(b_3^*, p_3^*) = (\$5, \$1.25)$.[15] Assignment was at the individual level, resulting in treatment variation within school, grades, and classrooms.

More specifically, 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 random assignment that minimized overall correlations between treatment status and balance variables. A balance table (Table 9) in the Online Supplemental Appendix presents correlations between gender, individual race groups, grade level, and pre-test score. This table verifies that our final treatment assignment was independent of all balancing variables. Although not reported in the table, treatment assignments were also independent of school district, by construction, as explained above.

Our pre-exam materials were produced and organized in such a way that they could be collected from teachers and rapidly processed so as to allow for balancing on initial math proficiency during randomization. Exams were administered to students toward the end of the school week, and they were processed, randomization executed, and personalized instruction materials for each student were produced over the weekend for in-classroom delivery by math teachers the following Monday. Each student participant received a personalized letter in a sealed envelope, containing login credentials, instructions for accessing the website, and their individual piece-rate incentive contract. They were also promised prompt delivery of payments within 2 weeks following the end of the experiment (which actually happened).

The structure of our incentives had several advantages that encouraged effort from the students so we could better infer the two central parameters of our model (θ_e, θ_l). First, we incentivized successful completion of learning tasks rather than the time spent on these tasks. This is consistent with actual school environments where students are typically rewarded or punished based on whether they complete assignments. Furthermore, we incentivized short-run tasks (analogous to a short homework assignment)

[15]Base payments varied inversely with marginal wage only to mitigate possible concerns of fairness on the part of participant households. A pilot study indicated an expected average output of ≈ 20 quizzes per student, at which point total payments across all three contracts are equal.

rather than long-term outcomes such as year-end grades, making the decisions faced by students in our sample more consistent with their frequent decisions day-to-day.

Second, we kept the window of effort short, in terms of both the size of incentivized tasks and in terms of payment delivery, to minimize the temporal gap between effort and reward. Our website continually reported total earnings increases each time a student passed a new quiz, and promised payments followed promptly after the post-exam. Incentives are more effective when rewards follow actions as soon as possible.[16] Third, we allow students multiple opportunities to attempt to pass each quiz. Thus, failed attempts can still motivate students to exert additional effort to achieve the intended result (Berger & Pope, 2011). High-frequency feedback on performance is also a key aspect of helping students learn about their own ability to convert time on task into academic achievement.

3.3.4. *Experiment Timeline.* In summary, the experiment took place as follows.

(1) Students took a pre-test and survey administered by their teachers in class.
(2) Students were randomly assigned a wage contract, and provided with information about the experiment, including the website and their earnings potential.
(3) For the next 10 days, student work on the website counted towards their compensation. Following the 10 day period, they were paid based on the number of quizzes successfully completed during that time.
(4) Students took a post-test and a second survey administered by their teachers in class.

3.4. **Descriptive Statistics.** Table 2 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" (URMs). This convention follows the higher education literature, where Blacks and Hispanics are known to be proportionally under-represented at post-secondary education institutions generally, and especially under-represented at elite colleges and universities. By contrast, Asian students, although a statistical demographic minority group, are proportionally over-represented at colleges generally, and particularly so at elite colleges, like their White counterparts. Thus, Asians do not satisfy the definition of a "URM" group. For ease of discussion, we will often refer to URMs as simply "minorities" for short, while recognizing this important caveat.

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 A), 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 extra-curricular activities. Whites/Asians have access to more adult academic helpers (including parents, grandparents, and tutors) and were more likely to be enrolled in sports and music. Black and Hispanic students are more likely to report that math is either their favorite or least favorite subject relative to their White/Asian peers. Minority students also self-reported

[16]Bettinger (2012) found evidence that incentives announced at the start of the year for performance on the end-of-year test have little impact, while Levitt, List, and Sadoff (2016) found that incentives offered immediately before students take a test have a large impact (and, likewise, delaying payment after the test can have large effects on effort). Minimizing temporal distance between the required effort and the delivered reward can be particularly helpful for groups that have high discount rates according to Bettinger and Slonim (2007).

TABLE 2. DESCRIPTIVE STATISTICS: SURVEY & SOCIOECONOMIC VARIABLES BY SUB-SAMPLE

SUB-SAMPLE:	**ALL**	**FEMALE**	**MALE**	**BLACK**	**HISPANIC**	**WHITE/ASIAN**
HEADCOUNT/FRACTION OF TOTAL:	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
(std. dev.)	*(41,470)*	*(41,107)*	*(41,871)*	(32,390)	*(23,175)*	*(24,602)*
Nbhd Uninsured Minor Rate	0.252	0.253	0.252	0.378	0.616	0.072
(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
(std. dev.)	*(0.499)*	*(0.500)*	*(0.499)*	*(0.081)*	*(0.205)*	*(0.364)*
District 2	0.268	0.260	0.276	0.650	0.103	0.136
(std. dev.)	*(0.443)*	*(0.439)*	*(0.447)*	*(0.478)*	*(0.304)*	*(0.343)*
District 3	0.267	0.266	0.269	0.344	0.854	0.021
(std. dev.)	*(0.443)*	*(0.442)*	*(0.444)*	*(0.475)*	*(0.354)*	*(0.144)*
FAMILY & RECREATIONAL TIME-USE VARIABLES						
# Adult Academic Helpers	1.140	1.163	1.117	1.128	0.615	1.328
(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
(std. dev.)	*(0.783)*	*(0.792)*	*(0.756)*	*(0.825)*	*(0.766)*	*(0.765)*
# Video Gaming Systems at Home	1.570	1.474	1.660	1.648	1.480	1.554
(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
(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
(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
(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
(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
(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
(std. dev.)	*(0.172)*	*(0.172)*	*(0.171)*	*(0.167)*	*(0.181)*	*(0.158)*
ACADEMIC PREFERENCES & ATTITUDE VARIABLES						
Math is Favorite Subject	0.361	0.319	0.404	0.431	0.439	0.302
(std. dev.)	*(0.480)*	*(0.466)*	*(0.491)*	*(0.496)*	*(0.497)*	*(0.460)*
Math is Least Favorite Subject	0.216	0.254	0.176	0.277	0.212	0.189
(std. dev.)	*(0.411)*	*(0.435)*	*(0.381)*	*(0.448)*	*(0.410)*	*(0.392)*
Extrinsic Motiv. Score (standardized)	0	-0.023	0.024	-0.222	-0.030	0.122
(std. dev.)	*(1)*	*(0.989)*	*(1.011)*	*(1.016)*	*(1.005)*	*(0.971)*
Intrinsic Motiv. Score (standardized)	0	0.056	-0.058	0.010	0.150	-0.059
(std. dev.)	*(1)*	*(1.005)*	*(0.992)*	*(1.047)*	*(1.057)*	*(0.949)*
EXAM SCORES						
Pre-Test Score	13.40	12.71	14.11	7.93	7.94	18.07
(std. dev.)	*(8.96)*	*(8.23)*	*(9.62)*	*(6.13)*	*(6.10)*	*(8.35)*
Change in Score (Post-Pre)	1.55	1.94	1.14	0.88	0.49	2.20
(std. dev.)	*(5.00)*	*(5.03)*	*(4.94)*	*(5.01)*	*(4.89)*	*(4.94)*

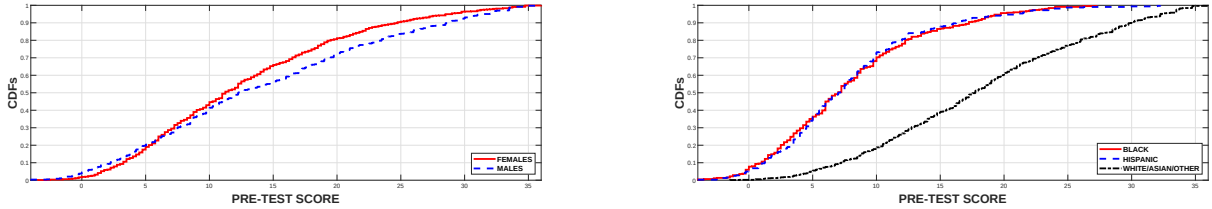
Notes: **Adult Academic Helpers** included parents, grandparents, and tutors; **Peer Academic Helpers** included siblings and friends. Numbers reported for Neighborhood Mean Income represent the median across all students in the sample. All other figures represent sample means, with sample standard deviations in parentheses and italics. Fifth-graders make up 47.3% of the total sample, with 6th-graders comprising the other 52.7%. Sub-sample proportions are close to that ratio for all gender and race groups.

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

Finally, Table 2 shows average pre-test scores by sub-group. The average male student correctly answered 1.4 additional questions on the assessment compared to the average female student. This corresponds to 0.16 SD higher score for males. The gender gap is relatively small compared to racial gaps in scores. White/Asian students performed substantially higher on the standardized mathematics

[17]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 1. Mathematics Pre-Test Scores by Gender and Race



pre-assessment than their Black and Hispanic peers, with the average White/Asian student correctly answering more than 10 additional questions, as compared to the mean minority student. This is roughly a 1.13 SD higher score for the mean White/Asian student.

Figure 1 illustrates the pre-test score distributions by gender and race. In the left panel we observe that low-achieving females slightly outperform low-achieving males (approximately the lowest quartile). Among high achievers the gender gap favors males, with a more-substantial gap among those who perform above average on the pre-test. There is little difference in the initial proficiency distributions of Black and Hispanic students (right panel), but there is a substantial gap between them and their White/Asian peers. These observed performance gaps in our pre-test scores collected during the experiment are consistent with evidence of substantial demographic achievement gaps from other studies (e.g. Clotfelter et al., 2009; Hanushek & Rivkin, 2006, 2009; NAEP, 2019).

Finally, Table 3 displays descriptive statistics of students' logged activities on our math website. Moving forward it will be helpful to define "workers" as the group of students who completed at least $Q_i \geq 2$ modules on the math website, and "non-workers" as students who did not. Workers constituted roughly half of the sample population (see Figure 2 below), though it is important to keep in mind that selection into worker status is a function of both θ_e and θ_l. The top panel of Table 3 pertains to all students, and the middle panel to workers only. The table depicts considerable raw differences across students in terms of willingness to spend time on math, rate of progress, and volume of learning tasks completed. Half of students logged no time on the website, while 4% of them completed all 80 learning modules. The highly skewed distributions of different measures have medians all being well below the means, and standard deviations generally being near or well above the means.

To place these figures in context, first note that website activity was above and beyond a child's regular schoolwork regimen. For a basis of comparison, we compiled data on school homework time per-day in our pre-survey and post-survey.[18] Importantly, the daily homework measure covers time spent on *all* school subjects, not just math. One possible threat to our identification strategy would be if students responded to our financial incentives by neglecting their schoolwork in proportion to the strength of the incentives offered. However, in multiple conversations with administrators and teachers they universally reported back to us a firm impression that the kids displayed no change in how much homework they were actually turning in during our study period. Our survey data appear to corroborate this claim: for

[18]To obtain this information, 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)?". To make it easy for children to think about the appropriate answer to this question, available responses were multiple choice: "a. None; b. Less than one hour per day; c. Between one hour and two hours per day; d. Between two and three hours per day; e. More than three hours per day," and we coded a.– e. as 0, 1, 2, 3, and 4 hours, respectively. We repeated both questions on the post-survey as well, but there we asked students to think about the previous two weeks, specifically. We then averaged across responses on the pre- and post-surveys. Finally, for a child's average daily time spent, we used the formula $(4/7) \times$ *weekday avg. daily homework time* $+ (3/7) \times$ *weekend avg. daily homework time*.

TABLE 3. DESCRIPTIVE STATISTICS: WEBSITE ACTIVITY & DAILY HOMEWORK TIME

Variable	Mean	Median	Std. Dev.	N	Contract Group 1 Mean	Contract Group 2 Mean	Contract Group 3 Mean
WEBSITE ACTIVITY: ALL STUDENTS							
Quizzes Passed	10.04	1	19.64	1,676	7.55	10.41	12.16
Math Problems Solved	60.25	6	117.86	1,676	45.29	62.47	72.93
Website Time Logged (minutes)	82.63	16.79	154.48	1,676	61.89	82.55	103.34
Total Pay	$14.77	$0.00	$23.80	1,676	$11.94	$14.89	$17.46
WEBSITE ACTIVITY: "WORKERS" ONLY							
Quizzes Passed	22.34	12	24.29	749	17.72	22.91	25.96
Math Problems Solved	134.07	72	145.73	749	106.31	137.45	155.77
Website Time Logged (minutes)	176.61	109.66	192.26	749	135.81	176.01	213.93
Within-Child Time Per Passed Quiz (minutes)	11.11	8.08	9.62	749	10.47	11.32	11.49
Total Pay	$33.04	$21.75	$25.77	749	$28.29	$32.91	$37.45
SELF-REPORTED AVG. DAILY HOMEWORK TIME ACROSS ALL ACADEMIC SUBJECTS							
All Students (hours)	1.248	1.214	0.681	1,676	—	—	—
Workers Only (hours)	1.424	1.429	0.647	749	—	—	—

Notes: "Workers" are the set of all students who passed at least 2 quizzes on the website and received a positive payout.

the sub-sample of worker students the homework time reports across the pre- and post-survey differed on average by a small (≈1%) and statistically insignificant amount ($p-value=0.765$).

Aside from providing a robustness check, this result allows us to use daily homework time as a useful benchmark for judging the magnitude of logged website activity. If we assume that mathematics accounted for between 25% and 50% of daily homework time, then among workers the average (median) website math time would have represented an increase of between 41% and 83% (26% and 51%), relative to regular math homework. Of course, this number would understate the magnitude of learning task volume increase relative to the average student, since non-workers have systematically lower academic efficiencies. This can be seen in that mean time per passed quiz trends upward between contract groups 1, 2, and 3: as offered piece-rate incentives increase, a marginal group of students having higher θ_e's self-selects into the worker group.

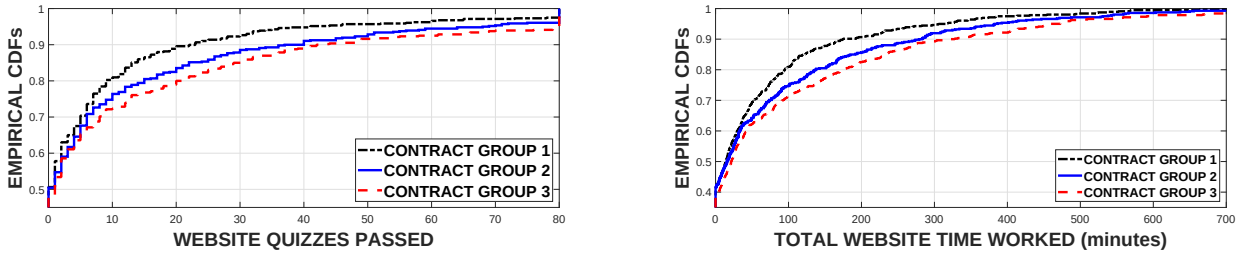
For an alternate benchmark of math work volume, we discussed the figures on website activity for workers (middle panel) with a mathematics education consultant employed by a state board of education for a mid-western U.S. state. Although volume of math problems assigned varies from classroom to classroom, the consultant expressed the opinion that 72 extra math problems solved within a 10-day period (the median for the worker group) would be an increase of between 50% and 100% in terms of regularly assigned homework volume for an average 5^{th}- or 6^{th}-grade student. Thus, overall we see that for some students our incentives induced substantial increases of learning task volume, though the distribution of the increase is heavily skewed.

4. ESTIMATION METHODOLOGY

4.1. **Student Time Allocation Model.** Estimation of student time allocation concentrates on quantifying three model primitives: individual-level academic efficiency, θ_{ei}, individual-level time preference, θ_{li}, and the common cost function $c(t)$. Along the way we also estimate several parameters of secondary interest, such as τ_0, γ, and the distributions of work-time shocks.

4.1.1. *Academic Efficiency and Work-Time Shock Distributions.* Estimation of the academic efficiency parameter hinges on panel-data methods using the within-child series of observed work times, $\left\{\{\tau_{q_i}\}_{q_i=1}^{Q_i}\right\}_{i=1}$.

FIGURE 2. Math Website Output and Work Time by Contract Group



Taking logs of both sides of equation (2) provides the following equality

$$\log(\tau_{q_i}) = \log(\tau_0) + \log(\theta_{ei}) - \gamma \log(q_i) + \log(u_{q_i}), \quad q_i = 1, \ldots, Q_i, \quad \{i | i = 1, \ldots, N, \quad Q_i \geq 0\}.$$

This constitutes a linear-in-parameters regression equation where individual heterogeneity enters as a student fixed-effect and τ_0 and γ, serve as an intercept and slope term. We estimate regression parameters through a standard differencing approach.[19] A key complication is that student fixed effects can only be inferred for the set of workers. This issue will be addressed in the next two sections.

Using regression estimates we can back out the distribution of production time shocks from the fitted residuals $\widehat{u}_{q_i} = \tau_{q_i} / (\widehat{\tau}_0 \widehat{\theta}_{ei} q_i^{-\widehat{\gamma}})$. We allow for heteroskedastic shocks by partitioning the support of $\widehat{\theta}_{ei}$ into 5 sub-intervals of equal length $I_j = \left[\min(\widehat{\theta}_{ei}) + (j-1)h, \min(\widehat{\theta}_{ei}) + jh\right]$, where $h = (\max(\widehat{\theta}_{ei}) - \min(\widehat{\theta}_{ei}))/5$ is the length of each interval $j = 1, \ldots, 5$.[20] Then, we estimate conditional shock distributions by splitting the sample of fitted residuals into 5 sub-samples $\left\{\widehat{u}_{q_i} | \widehat{\theta}_{ei} \in I_j\right\}$, and smoothing the corresponding empirical CDFs using flexible B-splines $F_u(u | \theta_{ei} \in I_j; \boldsymbol{\pi}_{uj})$ with parameters $\boldsymbol{\pi}_{uj}$.[21]

4.1.2. *Labor-Supply.* We estimate the time preference trait and labor-supply cost function through a simulated GMM approach. The identification framework proposed by Torgovitsky (2015) and D'Haultfoeuille and Février (2015) relies on discrete instruments to create shifts in observable distributions of incentivized actions across groups of agents that are otherwise identical in their distributions of unobservables. Figure 2 demonstrates that these conditions are satisfied by our field experimental controls: contract variation induced stochastic dominance shifts in the CDFs of T and Q, while individual-level randomization ensures that students receiving those contracts are otherwise the same. Under these conditions, Torgovitsky (2015) and D'Haultfoeuille and Février (2015) show that counterfactual comparisons across similar agents working under different incentives are enough to uniquely disentangle the shape of the common utility function from idiosyncratic agent-level heterogeneity. Our simulated GMM estimator is explicitly built upon functional representations of these counterfactual comparisons. In order to facilitate this undertaking we start with a flexible, parametric, B-Spline specification of the common cost function $c(t; \boldsymbol{\pi}_c)$, having parameter vector $\boldsymbol{\pi}_c$ (to be estimated) which uniquely determines its shape.[22]

[19] Note that the $\widehat{\theta}_{ei}$ estimates have differing variances due to the unbalanced panel (i.e., Q_i varies across students).

[20] We also tried a finer partition of 10 sub-intervals of the support of θ_e, but it made little difference in the following stage of estimation, relative to the specification with 5 sub-intervals, while increasing computational requirements.

[21] We chose 4 knots, uniformly placed in quantile rank space. After constraining the endpoints—a CDF must equal 0 and 1 at the extremes of the support—this left 5 free parameters, $(\pi_{u2j}, \ldots, \pi_{u6j})$, to fit the empirical CDFs of residuals. The tight fit between the two is depicted in Figure 16 in the online supplemental appendix.

[22] For the common cost function we chose 6 knots, placed evenly at the quintiles and endpoints of the sample of time worked. We then added three extra knots, uniformly spaced in the upper quintile to target extra flexibility and deal with a long, skewed upper tail. After imposing the two boundary conditions in Assumption 3, this left 10 free parameters to allow the model-generated CDFs of Q_i to fit their empirical analogs.

With this parametric form, individual choices of T_i and Q_i are the basis for inference on the shape of the cost function, which in turn uniquely determines each individual's time preference parameter θ_{li}.

To fix ideas, consider individual i, whose quiz output Q_i was at quantile rank r_i in contract group 1. Holding fixed the cost function parameters $\boldsymbol{\pi}_c$, we can reverse-engineer time preference by repeatedly simulating sequences of work times using θ_{ei} and $F_u(u|\theta_{ei})$. We choose the value of θ_{li} such that, given her actual assignment to contract (b_1^*, p_1^*), optimal stopping choices in Q space (under decision problem (4)) imply mean production time across all simulated outputs equal to i's observed choice T_i. Moreover, i's choice of work volume Q_i can inform us about the shape of the cost function. Specifically, Q_i first contributes to the empirical CDF of work volume under assignment to contract 1:

$$\widehat{F}_q(q|b_1^*, p_1^*) = \sum_{j=1}^{N} \mathbb{1}\frac{\left[Q_j \leq q \;\&\; p_j = p_1^*\right]}{\sum_{j=1}^{N} \mathbb{1}\left[p_j = p_1^*\right]}.$$

Second, with the value of θ_{li} known, we can further simulate a sequence of *counterfactual* work volume choices $\{Q_{2is}^*\}_{s=1}^S$ under contract 2, and $\{Q_{3is}^*\}_{s=1}^S$ under contract 3. These simulated values depend on θ_{ei} and $F_u(u|\theta_{ei})$, which are both fixed at this stage, and on the shape of the cost function $c(\cdot;\boldsymbol{\pi}_c)$ (which also determines θ_{li}). They contribute to the model-generated CDFs of work volume under assignment to contracts 2 and 3 through the following relationship:

$$\widetilde{F}_q(q|b_k^*, p_k^*; \boldsymbol{\pi}_c) = \sum_{j=1}^{N}\sum_{s=1}^{S} \mathbb{1}\frac{\left[Q_{kjs}^* \leq q \;\&\; p_j \neq p_k^*\right]}{\sum_{j=1}^{N} \mathbb{1}\left[p_j \neq p_k^*\right] \times S}, \quad k = 2, 3.$$

Thus, child i's observed choices contribute to the empirical CDF for her actual contract assignment 1, and they also contribute indirectly (by determining θ_{li}) to the model-generated CDFs under counterfactual contract assignments 2 and 3. Intuitively, cost function parameters $\boldsymbol{\pi}_c$ are then chosen to match child i's counterfactual projections to those of children at quantile rank r_i in contract groups 2 and 3.

Although this is the basic intuitive form of the GMM estimator, there are two complications regarding mass points at the extremes of the sample. First, we have a small mass of students who achieve full output $Q_i = 80$ on the website, as can be seen in Figure 2. This means that their academic efficiency trait, θ_{ei}, is known, but without extra structure their time preference trait, θ_{li}, can only be bounded from above. This is because it is impossible to know whether a given individual would have optimally chosen *exactly* $Q_i = 80$, or $Q_i > 80$ if given the chance. We deal with this problem by estimating a constrained quantile function using a low-dimensional B-spline to extrapolate into the missing upper tails of the empirical CDFs of Q. After discretizing the upper tail (for computational tractability), for each individual with full output this renders up to 5 possibilities for optimal stopping points $\{\widehat{Q}_{i1}, \ldots, \widehat{Q}_{i5}\}$, all being at or above 80.[23] For each $(\theta_{ei}, \widehat{Q}_{im})$ pair, $m = 1, \ldots, 5$, we back out a time preference trait $\theta_{li}(\widehat{Q}_{im})$ to match $\widehat{Q}_{im}$ as the optimal stopping point, and we run counterfactual simulations for each $(\theta_{ei}, \theta_{li}(\widehat{Q}_{im}))$ pair. However, we give each of these $(1/5)^{\text{th}}$ weight when incorporating them into the model-generated CDFs $\widetilde{F}_q$.

The second and more challenging mass-point problem pertains to the sizeable fraction of students who chose not to complete the minimum output for pay: those with $Q_i < 2$. For these individuals we can

[23]The extrapolating B-spline quantile functions overlapped their empirical counterparts to the 85^{th} percentile. We assumed that no student would choose to more than double the available workload on the website, so tails were bounded from above by Q=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. We discretized the extrapolated tails 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 $Q_i = 80$. 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 17 in the online supplement plots the extrapolated upper tails against the empirical CDFs of Q.

only infer that their 2-dimensional traits are within a contract-specific region bounded by a decreasing function $\underline{\Theta}_l(\theta_e; b_k^*, p_k^*, \boldsymbol{\pi}_c)$ for $k = 1, 2, 3$, where to the northwest of this boundary either θ_{ei} was too high or θ_{li} was too high (or both) to rationalize positive output in response to their contract (b_k^*, p_k^*). For most of these individuals, their counterfactual outputs under alternate contract assignments would likely be zero as well. However, some fraction of them may be marginal agents where counterfactual assignment to one of the alternative contracts k' might induce a change to positive output. This we must correct for when computing the simulated CDFs $\widetilde{F}_q$. To do so, we use our distributional Assumption 1 to integrate over the non-identified portion of the space for counterfactual output simulations.

Essentially, this procedure is a 2-dimensional variant of Heckman's (1979) classic sample-selection correction, where the selection locus $\underline{\Theta}_l(\theta_e; b_k^*, p_k^*, \boldsymbol{\pi}_c)$ is known. More concretely, holding $\boldsymbol{\pi}_c$ fixed, all $(\theta_{ei}, \theta_{li})$ pairs can be inferred for students with $Q_i \geq 2$, and the upper bound $\underline{\Theta}_l(\theta_e; b_k^*, p_k^*, \boldsymbol{\pi}_c)$ on the identified set can be computed as the northwest boundary of the convex hull of the set $\{(\theta_{ei}, \theta_{li}) | Q_i \geq 2,\ p_i = p_k^*\}$. Next, the parameters of the bivariate log-normal distribution, $(\theta_e, \theta_l) \sim BVlN(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$, are pinned down by matching the selection frequency as well as the selected means, variances, and covariance of (θ_e, θ_l), conditional on $Q \geq 2$, which adds some additional moment conditions to the GMM objective function.[24] For contract k, we denote the selected empirical moments by

$$\widehat{M}(\boldsymbol{\pi}_c, k) = \left[\widehat{P}(k),\ \widehat{E}_e^1(\boldsymbol{\pi}_c, k),\ \widehat{E}_l^1(\boldsymbol{\pi}_c, k),\ \widehat{E}_e^2(\boldsymbol{\pi}_c, k),\ \widehat{E}_l^2(\boldsymbol{\pi}_c, k),\ \widehat{E}_{el}^3(\boldsymbol{\pi}_c, k)\right]^\top,$$

$$\textit{(selection frequency)}\ \ \widehat{P}(k) = \frac{\sum_{i=1}^N \mathbb{1}[Q_i \geq 2\ \&\ p_i = p_k^*]}{\sum_{i=1}^N \mathbb{1}[p_i = p_k^*]},$$

$$\textit{(selected raw moments)}\ \ \widehat{E}_j^r(\boldsymbol{\pi}_c, k) = \frac{\sum_{i=1}^N \log(\theta_{ji})^r \mathbb{1}[Q_i \geq 2\ \&\ p_i = p_k^*]}{\sum_{i=1}^N \mathbb{1}[p_i = p_k^*]},\ j = e, l,\ r = 1, 2,$$

$$\textit{(selected product moment)}\ \ \widehat{E}_{el}^3(\boldsymbol{\pi}_c, k) = \frac{\sum_{i=1}^N \log(\theta_{ei}) \times \log(\theta_{li}) \mathbb{1}[Q_i \geq 2\ \&\ p_i = p_k^*]}{\sum_{i=1}^N \mathbb{1}[p_i = p_k^*]},$$

and we denote their model-generated analogs by

$$\widetilde{M}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k) = \left[\widetilde{P}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k), \widetilde{E}_e^1(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k), \widetilde{E}_l^1(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k), \widetilde{E}_e^2(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k), \widetilde{E}_l^2(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k), \widetilde{E}_{el}^3(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k)\right]^\top.$$

These last moments are determined by computing the analogous integrals (using the bivariate log-normal density) over the selected-in region. This is the reason for the dependence on the cost function parameters $\boldsymbol{\pi}_c$ (through their influence on the selection thresholds).

Finally, for computational tractability we perform stochastic integration when computing $\widetilde{M}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k)$. We also perform stochastic integration over the non-identified region (i.e., to the northwest of the selection loci $\underline{\Theta}_l(\theta_e; b_k^*, p_k^*, \boldsymbol{\pi}_c),\ k = 1, 2, 3$) when simulating counterfactual choices for individuals who chose $Q_i < 2$ under their actual contract assignment. For stochastic integration, we simulate a sample of independent standard normal draws, $\boldsymbol{Z} = [Z_e, Z_l]$, where $Z_m = [z_{m1}, \ldots, z_{mT}]^\top$ and $m = e, l$. At each iteration of the solver, these can be transformed into bivariate log-normal random variables through

$$(\boldsymbol{\theta}_e, \boldsymbol{\theta}_l) = \exp\left(\boldsymbol{V}\boldsymbol{Z} + \widetilde{\boldsymbol{\mu}}\right), \tag{9}$$

where $\boldsymbol{V}$ is the lower-triangular component of the Cholesky decomposition of the covariance matrix $\widetilde{\boldsymbol{\Sigma}}$. Finally, for each $k = 1, 2, 3$ we discard all resulting (θ_e, θ_l) pairs to the southwest of the selection locus for contract k, and for each remaining pair we repeatedly simulate optimal counterfactual choices

[24] Note that the bivariate log-normal parameters mentioned here, $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$, are different from those referenced in Assumption 1, where the means are zero and $\boldsymbol{\Sigma}$ is the covariance matrix of the idiosyncratic components (η_{ei}, η_{li}).

under the other two contracts, as was done for other students. This sample of simulated choices under counterfactual contract k' is then appropriately scaled when computing $\widetilde{F}_q(q|b^*_{k'}, p^*_{k'}; \boldsymbol{\pi}_c)$ according to the mass of contract-k students who chose $Q_i = 0$.

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

$$\begin{aligned}\Big[\widehat{\boldsymbol{\pi}}_c, \widehat{\widetilde{\boldsymbol{\mu}}}, \widehat{\widetilde{\boldsymbol{\Sigma}}}\Big] = \text{argmin} \Bigg\{ & \rho_0 \sum_{q=2}^{80} \sum_{k=1}^{3} \omega_c^k(q) \Big(\widehat{F}_q(q|b^*_k, p^*_k) - \widetilde{F}_q(q|b^*_k, p^*_k; \boldsymbol{\pi}_c) \Big)^2 \\ & + \sum_{k=1}^{3} \Big(\widehat{M}(\boldsymbol{\pi}_c, k) - \widetilde{M}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k) \Big)^\top \boldsymbol{\rho} \Big(\widehat{M}(\boldsymbol{\pi}_c, k) - \widetilde{M}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \boldsymbol{\pi}_c, k) \Big) \Bigg\} \\ & s.t. \quad c(0; \boldsymbol{\pi}_c) = 0, \;\; c'(0; \boldsymbol{\pi}_c) = 1,\end{aligned} \tag{10}$$

Some final comments on implementation are in order. First, we used an inverse-variance weighting scheme $\omega_c^k(q) \equiv \widehat{F}_q(q|b^*_k, p^*_k)(1 - \widehat{F}_q(q|b^*_k, p^*_k))$, $k = 1, 2, 3$, that places more emphasis on matching segments of the empirical CDFs that are more precisely estimated. Second, we implemented our GMM estimator using the *mathematical programming with equilibrium constraints, or MPEC* approach pioneered in the economics literature by Su and Judd (2012). This proved to be much faster and numerically stable than the alternative *nested fixed-point* approach, which would require serially optimizing the second set of moments in equation (10) for each iteration of the cost function parameter vector. Instead, the MPEC approach allows both $\boldsymbol{\pi}_c$ and $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ to update independently along the path to convergence, at which point both sets of moment conditions are mutually optimized. The purpose of the penalty parameters ρ_0 and $\boldsymbol{\rho}$ is to ensure that both sets of moment conditions are roughly on the same order of magnitude, and that sufficient attention is paid to crucial aspects of the selection equations.[25]

4.2. **Decomposition of Student Characteristics.** We now turn to the decomposition of student traits into a predictable component and an idiosyncratic component:

$$\log(\theta_{ei}) = \boldsymbol{X}_{ei}\boldsymbol{\beta}_e + \eta_{ei}, \; i = 1, \ldots, N, \tag{11}$$

$$\log(\theta_{li}) = \boldsymbol{X}_{li}\boldsymbol{\beta}_l + \eta_{li}, \; i = 1, \ldots, N. \tag{12}$$

The covariate vector, $\boldsymbol{X}_{ei}$, for the academic efficiency 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) and *fraction of minors with no private health insurance* (a proxy for deprivation of non-school developmental resources). The covariate vector $\boldsymbol{X}_{li}$ for time preferences 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*; and *weekday time spent on recreational internet use*. The idea in adding these additional factors to equation (12) is that θ_{li} 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

[25] We set ρ_0=100 so that the primary moments are on the same order of magnitude as the selection moments, and $\boldsymbol{\rho}_{\{1,1\}}$=10, $\boldsymbol{\rho}_{\{i,j\}}$=1, i=j>1, and $\boldsymbol{\rho}_{\{i,j\}}$=0, $i \neq j$, i=1,...,6, $j = 1, \ldots, 6$ in order to place particular emphasis on matching the empirical selection frequency.

or her responsiveness to different forms of incentives, holding her academic efficiency θ_{ei} fixed. These variables are all summarized in Table 2.

The challenge here is a basic sample truncation problem: while $(\boldsymbol{X}_{ei}, \boldsymbol{X}_{li})$ is known for all $i = 1, \ldots, N$, the outcome variables $(\log(\theta_{ei}), \log(\theta_{li}))$ are known only for students who chose $Q_i \geq 2$. By adopting Assumption 1, $(\eta_e, \eta_l) \sim BVN(\mathbf{0}, \boldsymbol{\Sigma})$, we can implement a 2-dimensional Maximum Likelihood Tobit strategy, using the known, contract-specific selection thresholds $\underline{\Theta}_l(\theta_e; b_k^*, p_k^*, \widehat{\boldsymbol{\pi}}_c)$, $k = 1, 2, 3$, uncovered in the previous stage of estimation. Moreover, we allow for our covariance structure to depend on race and gender by adopting the following specification for $\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{ei}^2 & \sigma_{eli} \\ \sigma_{eli} & \sigma_{li}^2 \end{bmatrix}$:

$$\begin{aligned} \sigma_{ei} =& \sigma_{ei0} + \sigma_{ei1} fem_i + \sigma_{ei2} black_i + \sigma_{ei3} hispanic_i \\ \sigma_{li} =& \sigma_{li0} + \sigma_{li1} fem_i + \sigma_{li2} black_i + \sigma_{li3} hispanic_i \\ \sigma_{eli} =& \sigma_{eli0} + \sigma_{eli1} fem_i + \sigma_{eli2} black_i + \sigma_{eli3} hispanic_i. \end{aligned}$$

Our Tobit estimator is thus defined by optimizing the following log-likelihood function:

$$\begin{aligned} \left[\widehat{\boldsymbol{\beta}}_e, \widehat{\boldsymbol{\beta}}_l, \widehat{\boldsymbol{\Sigma}}\right] = \text{argmax} \Bigg\{ & \sum_{i=1}^{N} \mathbb{1}(Q_i \geq 2) \omega_{di} \log\left(f_{\eta_e, \eta_l}(\boldsymbol{X}_{ei}\boldsymbol{\beta}_e, \boldsymbol{X}_{ei}\boldsymbol{\beta}_e; \boldsymbol{\Sigma}_i)\right) \\ & + \mathbb{1}(Q_i < 2) \omega_{di} \log\left(\Pr\left[\log(\theta_l) > \log\left[\underline{\Theta}_l(\theta_e; b_i, p_i, \widehat{\boldsymbol{\pi}}_c)\right] \middle| \boldsymbol{X}_{ei}, \boldsymbol{X}_{li}; \boldsymbol{\beta}_e, \boldsymbol{\beta}_l, \boldsymbol{\Sigma}_i\right]\right) \Bigg\}, \end{aligned} \tag{13}$$

where the ω_{di} terms are inverse-variance weights: $\omega_{di} = \frac{1/Var(\widehat{\theta}_{ei}) + 1/Var(\widehat{\theta}_{li})}{2}$ whenever $Q_i \geq 2$, and $\omega_{di} = \min\{\omega_{dj} | Q_j \geq 2\}$ whenever $Q_i < 2$. For computational tractability, we compute the probability in the Tobit term above by simulation, similarly as we did above (see equation (9)).

4.3. **Skill formation Models.** The final stage of our empirical analysis is the estimation of the skill formation technology. For initial math skill, taking logs of both sides of equation (5) renders the following:

$$\log(S_i) = \boldsymbol{W}_i \boldsymbol{\alpha}_0 + \theta_{ei} \boldsymbol{W}_i \boldsymbol{\alpha}_e + \theta_{li} \boldsymbol{W}_i \boldsymbol{\alpha}_l + \log(\epsilon_i). \tag{14}$$

For the production technology of gains in math skill, we can re-write equation (7) as:

$$\Delta S_i = \boldsymbol{V}_i \boldsymbol{\delta}_0 + T_i \boldsymbol{V}_i \boldsymbol{\delta}_1 + T_i^2 \boldsymbol{V}_i \boldsymbol{\delta}_2 + Q_i \boldsymbol{V}_i \boldsymbol{\delta}_3 + Q_i^2 \boldsymbol{V}_i \boldsymbol{\delta}_4 + (T_i \times Q_i) \boldsymbol{V}_i \boldsymbol{\delta}_5 + \varepsilon_i, \tag{15}$$

where $\boldsymbol{V}_i = [\boldsymbol{W}_i, S_i, \theta_{ei}, \theta_{li}]$.[26] 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 household income* (a proxy for affluence) and *fraction of 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 both in the model of initial math skill and in the model of incremental skill gains, each of these factors is allowed to have a direct impact (through the intercept terms) and also to have an indirect impact (through the slope terms) of altering the map between the principal inputs and the final outputs.

While it has long been known that students attending schools with greater resources produce better outcomes (e.g., standardized test scores), it is unclear whether this is due to better school inputs *per se*, or whether it is attributable to selection of more academically adept students into those higher-performing schools. In short, to what extent are higher performing schools truly adding value versus

[26] 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.

merely shepherding gifted students through the academic pipeline? The figures in Table 1 suggest that making this distinction is far from obvious. In terms of studying the role of school quality in skill formation technology, a major advantage of our research design is that it first quantifies unobserved student traits, θ_e and θ_l, and thereby solves the classic endogeneity problem of omitted variable bias. The assumption that we require to attach a causal interpretation to estimates of parameters in the two production function equations above is the following:

Assumption 4. $E[\boldsymbol{W}_i^\top \log(\epsilon_i)|\theta_{ei}, \theta_{li}] = \mathbf{0}$ *and* $E[\boldsymbol{V}_i^\top \varepsilon_i|\theta_{ei}, \theta_{li}] = \mathbf{0}$.

There remain two final challenges to be addressed. First, since the empirical model of time allocation can only infer unique values of $(\theta_{ei}, \theta_{li})$ for students who chose minimal output $Q_i \geq 2$ on our website, we have a missing regressors problem in equations (14) and (15). This is fairly straightforward to address: using the Tobit maximum likelihood results from the previous section, for each student i with $Q_i<2$ we can compute the conditional expectations,[27]

$$\left(\widehat{\theta}_{ei}, \widehat{\theta}_{li}\right) = E\Big[\big(\log(\theta_e), \log(\theta_l)\big)\Big|\boldsymbol{X}_{ei},\ \boldsymbol{X}_{li},\ Q_i < 2,\ p_i;\ \widehat{\boldsymbol{\beta}}_e, \widehat{\boldsymbol{\beta}}_l, \widehat{\boldsymbol{\Sigma}}_i\Big].$$

The second challenge is that since student traits play the role of regressors in equations (14) and (15), sampling variability induces an errors-in-variables problem. To cope with this problem, we compute empirical Bayes (EB) estimates of (θ_e, θ_l). This approach reduces attenuation bias by shrinking fixed effect estimates toward their mean in proportion to the individual noise in each fixed effect. The 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 $\lambda_{ji} = \nu_j^2 / \left(\nu_j^2 + \nu_{rji}^2\right)$, where ν_j^2 is the estimated variance of true $\log(\theta_{ji})$, and ν_{rji}^2 is the estimated sampling residual variance on $\widehat{\log(\theta_{ji})}$ for individual i's trait $j = e,l$.[28] This results in the following EB estimates for student characteristics to be used as regressors for estimation of skill production technology:

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

Finally, the imputation of student traits for non-workers suggests that the error terms in equations (14) and (15) 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 by Wooldridge (2016).

4.4. **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 3.3.3) which balanced on race, gender, school district, grade level, and pre-test score. We begin by arranging all test subjects into race-gender-district-grade bins.[29] Suppose that there are K such bins in total, and that within contract $j = 1,2,3$

[27] This approach is in the spirit of standard methods for regression with X's surveyed by Little (1992, Section 4.2).

[28] 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 4.4.

[29] 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.

the bins each have $N_{1j},\ N_{2j}, \ldots, N_{Kj}$ subjects in them, respectively. Then, in order to construct a single block-bootstrap sample, for each bin, $k = 1, \ldots, K$, we do the following:

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

The final remaining question is how many bootstrap samples on which to generate and re-estimate the model. The main consideration here is a trade-off between simulation error and computational cost. Estimates of the student time allocation model generally took between 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, $\left\{\tau_0^{(s)},\ \gamma^{(s)},\ \boldsymbol{\pi}_c^{(s)}\right\}_{s=1}^{1,600}$, etc., with an individual's observables, $\left\{\{\tau_{q_i=1}^{Q_i}\}, T_i, Q_i, \boldsymbol{X}_{ei}, \boldsymbol{X}_{li}\right\}$, to compute bootstrapped fixed effect estimates $\left\{\theta_{ei}^{(s)}, \theta_{li}^{(s)}\right\}_{s=1}^{S}$. 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.

5. Empirical Results

5.1. **Cost Schedule, Time Preference, and Academic Efficiency Estimates.** Figure 3 illustrates the estimated cost function $C(T; \widehat{\theta}_l; \widehat{\boldsymbol{\pi}}_c)$ and marginal cost function, both scaled to the median value of θ_l among workers. The lower panel of the figure plots the histogram and density of total time worked T_i for context. Costs and marginal costs are precisely estimated for relatively low values of time expenditure, while the 95% confidence bands widen for higher values where the data are sparse. We find that a remarkably high degree of curvature in the common cost function $c\,(t; \widehat{\boldsymbol{\pi}}_c)$ is required to rationalize the observed distributions of work time and quiz outputs. The top panel of the figure labels cost levels at regular intervals to illustrate this point. Relative to a 90-minute time commitment, the depicted child's costs roughly quadruple with a doubling to 3 hours, and an additional doubling of time commitment slightly more than quadruples costs again (Figure 18 in the online appendix displays the goodness of fit that our flexible B-spline cost specification achieved). 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., students being offered *both* higher and lower incentives).

Figure 4 illustrates the degree of cost variation *across students*. The figure depicts cost schedules scaled to θ_l types at the 25$^{\text{th}}$ percentile, median, and 75$^{\text{th}}$ percentile of workers. The overall picture is one

FIGURE 3. Time Supply Cost & Marginal Cost Estimates

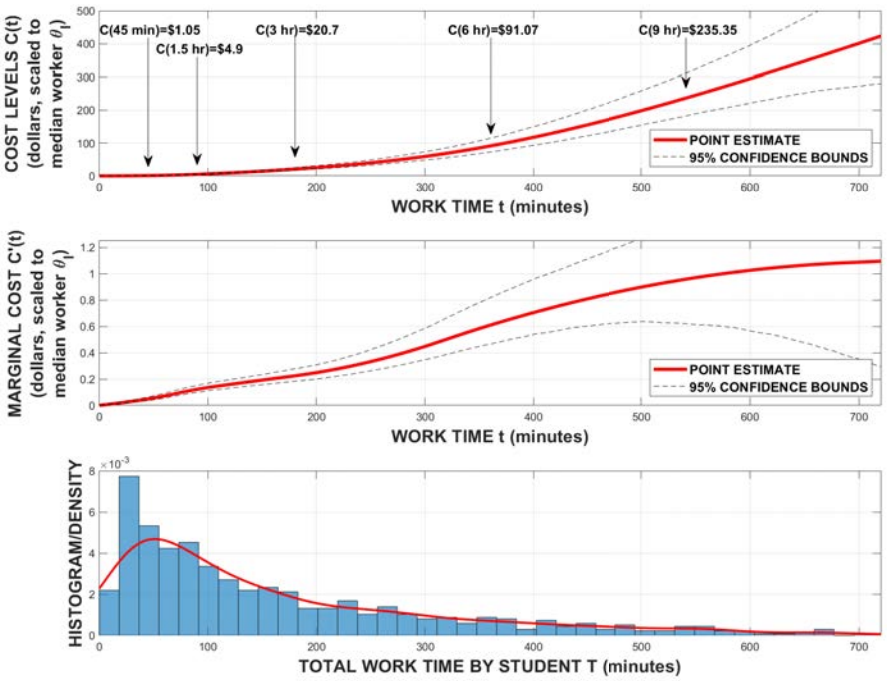
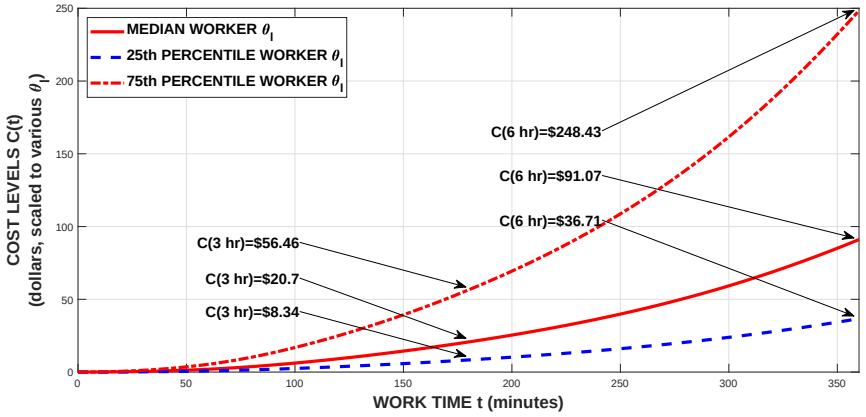


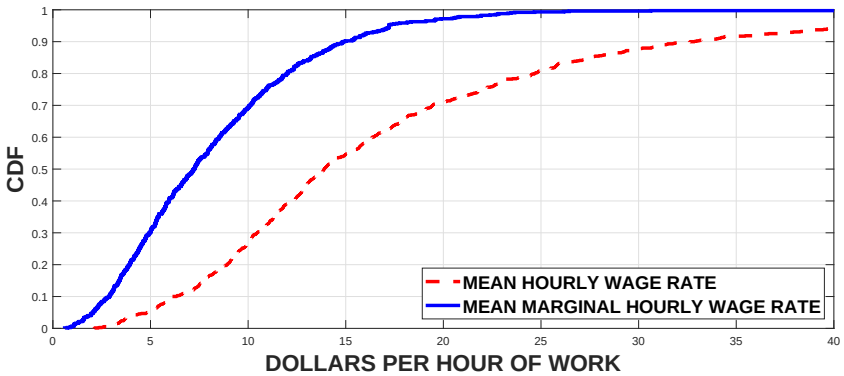
FIGURE 4. Time Supply Cost Function Estimates



of dramatic heterogeneity in willingness to supply time to math learning activity. Costs of 3-6 hours of foregone leisure differ by a factor of roughly 7 across the inter-.quartile range of worker types. Moreover, since the figure restricts attention to workers only, who have lower θ_l values, on average, the comparison across the 25th and 75th percentiles of the overall student distribution would be even more stark. Note, however, that Figures 3 and 4 consider costs of effort in the *time dimension only*.

As the model suggests, time costs θ_l do not determine a student's study effort choices alone; how productive they expect to be with their time, academic efficiency, plays a central role too. Figure 5 provides an intuitive illustration of heterogeneity across students in terms of θ_e. The figure plots two curves: the overall "mean hourly wage" gives the CDF of $(total\ payments\ to\ child\ i)/(total\ time\ worked\ by\ child\ i)$ and the "mean marginal hourly wage" is the CDF of $(total\ piece{-}rate\ payments\ to\ child\ i)/(total\ time\ worked\ by\ child\ i)$. This second measure is more conservative and ignores the fact that the first hour or so is most lucrative due

FIGURE 5. Effective Hourly Wage Rates



Notes: **Mean Hourly Wage Rate** is defined as $(total\ payments\ to\ child\ i)/(total\ time\ worked\ by\ child\ i)$. **Mean Marginal Hourly Wage Rate** is a more conservative measure which ignores base wage payments, or $(total\ piece-rate\ payments\ to\ child\ i)/(total\ time\ worked\ by\ child\ i)$.

TABLE 4. TOBIT REGRESSION RESULTS: ACADEMIC EFFICIENCY

SPECIFICATION:	**(1)**		**(2)**		**(3)**	
DEPENDENT VARIABLE: $\log(\theta_e)$	Estimate	St. Dev. Effect	Estimate	St. Dev. Effect	Estimate	St. Dev. Effect
Female $(\widehat{\beta}_1)$	0.2188***	**0.3278**	0.1760***	**0.2551**	0.1563***	**0.2411**
(std. err.)	*(0.0515)*		*(0.0499)*		*(0.0488)*	
Black $(\widehat{\beta}_2)$	0.7810***	**1.1702**	0.6248***	**0.9054**	0.5623***	**0.8677**
(std. err.)	*(0.1093)*		*(0.1185)*		*(0.1037)*	
Hispanic $(\widehat{\beta}_3)$	0.7873***	**1.1797**	0.4045**	**0.5863**	0.3655**	**0.5639**
(std. err.)	*(0.1710)*		*(0.1543)*		*(0.1364)*	
Grade 5 $(\widehat{\beta}_4)$	0.3096***	**0.438**	0.2940***	**0.4620**	0.2666***	**0.4114**
(std. err.)	*(0.0562)*		*(0.0514)*		*(0.0517)*	
District 2 $(\widehat{\beta}_5)$	—	—	0.2231**	**0.3234**	0.1333*	**0.2057**
(std. err.)			*(0.1038)*		*(0.0808)*	
District 3 $(\widehat{\beta}_6)$	—	—	0.7829***	**1.1346**	0.4919***	**0.7590**
(std. err.)			*(0.2031)*		*(0.1298)*	
Constant $(\widehat{\beta}_0)$	-0.3145***		-0.3926***		-0.3359***	
(std. err.)	*(0.0661)*		*(0.0856)*		*(0.0632)*	
Neighborhood SES Controls	yes		yes		yes	
Family Academic Support Controls	no		no		yes	
N	1,676		1,676		1,676	
Pseudo-R^2	0.378		0.397		0.370	
log-Likelihood	-3684		-3664.8		-3560.6	

Notes: *Higher* values of $\log(\theta_e)$ imply *lower* academic efficiency. **Neighborhood SES Controls** contain log of mean income and fraction of minors with no private health insurance. **Family Academic Support Controls** include (self-reported) counts of how many adults (e.g., parent, tutor, etc.), and how many peers (e.g., friend, sibling, etc.) regularly help the student with his/her math homework. In all model specifications, **Neighborhood SES Controls** individually play no statistically significant role, and **Academic Support Controls** individually play no economically significant role in explaining math academic efficiency. **St. Dev. Effect** represents the change in standard deviation units of $\log(\theta_e)$ from switching the value of a binary regressor from 0 to 1. Note that due to joint Tobit Estimation **Pseudo-R^2** for $\log(\theta_e)$ need not increase monotonically with model richness, though the sum of **Pseudo-R^2** for both $\log(\theta_e)$ and $\log(\theta_l)$ will generally rise.

to the one-time base wage payment. The median of "mean hourly wage" is \$13.98/hour and the median of "mean marginal hourly wage" is \$7.22/hour. Since test subjects are 9-11 years old, our offered piece-rate contracts translate into fairly strong incentives, on average, for children with academic efficiencies that are not too low. In fact, worker students above the 90th percentile (i.e., in the lower decile of θ_e) were making a comparable or better hourly wage to what many economics graduate students receive for teaching duties. Note that the majority of cross-student heterogeneity in Figure 5 derives from variation in academic efficiency, but once again, keep in mind that this figure does not incorporate a child's time preference θ_l. Our subsequent analyses will combine these two characteristics in various ways.

5.2. **Decompositions of Time Preferences and Academic Efficiency.** The most-substantial demographic differences in test scores are the racial differences between Black/ Hispanic and White/Asian students, driven by underlying differences in the distributions of θ_e across the demographic groups. Tables 4 and 5 report results from Tobit regressions exploring relationships between observable characteristics of students and their neighborhoods, and the type parameters we estimated using the structural model.

From Table 4, θ_e tends to be higher (i.e., lower academic efficiency) for females, Black students, and Hispanic students, compared to their male or White/Asian peers after conrolling for socioeconomic proxies. This means that males require less time to complete math learning activities, conditional on attempting them. Unsurprisingly, 6^{th}-grade students are more efficient than 5^{th}-grade students. In specification (1) females have average values of $\log(\theta_e)$ that are 0.33 SD below the values of their male peers, which is a little less than 3/4 of the gap in $\log(\theta_e)$ between 5^{th}- and 6^{th}-graders. Blacks and Hispanics tend to have values of $\log(\theta_e)$ that are 1.2 SD below their White/Asian peers, or approximately 2.7 times as large as an additional year of schooling. When we extend this analysis to control for a student's school district and their adult or peer support network, we still observe substantial differences due to gender and race. Females now lag males by 0.24 SD, which is 0.59 times the increase in $\log(\theta_e)$ due to an extra year of school. For the average Black (Hispanic) student in the sample $\log(\theta_e)$ tends to be 0.87 SD (0.56 SD) higher than the average White/Asian student, meaning their academic efficiency disadvantages are 2.1 (1.4) times the average effect of an additional year of schooling.

Alternatively, from Table 5 we observe that θ_l tends to be lower (i.e., higher motivation level for math) for females and Black students compared to their male and White/Asian peers. Hispanics also have lower mean θ_l, though the difference is not significant. Thus, on average females and minority students require fewer incentives to spend extra time working on math problems, compared to males and Whites/Asians. Similarly, 6^{th}-grade students require greater incentive than 5^{th}-graders to engage in extra math activity. This difference by grade, however, is relatively small compared to the differences due to gender or race. In model specification (1), females tend to have values of $\log(\theta_l)$ that are 0.59 SD below their male peers, and Black and Hispanic students tend to have values of $\log(\theta_l)$ that are 0.48 SD and 0.28 SD below their White/Asian peers, respectively. When we extend this analysis (specification (3)) to control for a student's school district, attitudes, preferences, time-use and consumption variables, and family/peer support network, we still observe substantial differences due to gender and race. The average female $\log(\theta_l)$ tends to be 0.42 SD lower than the average male $\log(\theta_l)$, and the average Black student $\log(\theta_l)$ tends to be 0.38 SD below that of the average White/Asian student. With these other controls, the effect for Hispanic students falls to only 0.05 SD below and remains insignificant.

Now we turn to the role of school quality in determining student ability and performance. In Table 4, even after controlling for observable student characteristics, attendance at a high-performing district induces lower values of θ_e. In other words, one's school enrollment predicts significant reductions in the time required for a student to complete learning tasks. Interestingly, from the descriptive evidence in Table 1, one might have suspected that District 1's inputs are more advantageous to the student than District 2's, which are in turn more advantageous than District 3's. 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 academic efficiency by 0.21 SD or 0.76 SD, respectively. The latter result is roughly 1.8 times the gap between grade-5 and grade 6-students, holding school district and all other student observables fixed.

TABLE 5. TOBIT REGRESSION RESULTS: TIME PREFERENCES

SPECIFICATION:	**(1)**		**(2)**		**(3)**	
DEPENDENT VARIABLE: $\log(\theta_l)$	Estimate	St. Dev. Effect	Estimate	St. Dev. Effect	Estimate	St. Dev. Effect
log(Mean Nbhd Income) (standardized) $(\widehat{\beta}_1)$	0.2148*	**0.1285**	0.0051	0.0032	0.0898	0.0534
(std. err.)	*(0.1430)*		*(0.1795)*		*(0.1072)*	
Nbhd Uninsured Minor Rate (standardized) $(\widehat{\beta}_2)$	0.8052***	**0.4855**	0.8061***	**0.5105**	0.3677***	**0.2204**
(std. err.)	*(0.1873)*		*(0.2416)*		*(0.1362)*	
Female $(\widehat{\beta}_3)$	-0.9766***	**-0.5882**	-0.8750***	**-0.5535**	-0.6953***	**-0.4163**
(std. err.)	*(0.1781)*		*(0.1588)*		*(0.1098)*	
Black $(\widehat{\beta}_4)$	-0.7971**	**-0.4801**	-0.5637	-0.3566	-0.6419**	**-0.3843**
(std. err.)	*(0.3413)*		*(0.3434)*		*(0.2620)*	
Hispanic $(\widehat{\beta}_5)$	-0.4691	-0.2826	-0.0842	-0.0532	-0.0862	-0.0516
(std. err.)	*(0.5038)*		*(0.5101)*		*(0.4682)*	
Grade 5 $(\widehat{\beta}_6)$	-0.3145**	**-0.1894**	-0.3186**	**-0.2015**	-0.2461***	**-0.1474**
(std. err.)	*(0.1389)*		*(0.1330)*		*(0.0936)*	
District 2 $(\widehat{\beta}_7)$	—	—	-0.4204	-0.2660	-0.0864	-0.0517
(std. err.)			*(0.2925)*		*(0.1732)*	
District 3 $(\widehat{\beta}_8)$	—	—	-1.1065	-0.7000	0.0037	0.0022
(std. err.)			*(0.7464)*		*(0.4119)*	
Math Favorite $(\widehat{\beta}_9)$	—	—	—	—	-0.2535***	**-0.1518**
(std. err.)					*(0.0934)*	
Math Least Favorite $(\widehat{\beta}_{10})$	—	—	—	—	0.0411	0.0246
(std. err.)					*(0.1550)*	
Extrinsic Motiv. Score $(\widehat{\beta}_{11})$	—	—	—	—	-0.6172***	**-0.3094**
(std. err.)					*(0.0881)*	
Intrinsic Motiv. Score $(\widehat{\beta}_{12})$	—	—	—	—	-0.5469***	**-0.2938**
(std. err.)					*(0.0735)*	
Constant $(\widehat{\beta}_0)$	-6.3750***	—	-6.2422***	—	-4.7812***	—
(std. err.)	*(0.2688)*		*(0.3475)*		*(0.4924)*	
Family Academic Support Controls	no		no		yes	
Extra-Curricular Controls	no		no		yes	
Gaming & Internet Use Controls	no		no		yes	
N	1,676		1,676		1,676	
Pseudo-R^2	0.061		0.076		0.206	
log-Likelihood	-3684		-3664.8		-3560.6	

Notes: *Higher* values of $\log(\theta_l)$ imply *higher* utility costs (*lower* willingness) of allocating time to extra math activity. The outcome variable $\log(\theta_l)$ represents a child's idiosyncratic willingness to substitute away from spending time on the outside option and toward extra study of mathematics. **Academic Support Controls** include self-reported tally of adults (e.g., parent, grandparent, tutor, etc.), and tally of peers (e.g., friend, sibling, etc.) that regularly help the student with his/her math homework. **Extra-Curricular Controls** (dummies for enrollment in sports, music, and clubs; and fraction of social time in structured, adult-supervised activities) individually do not play a statistically significant role in explaining leisure preferences. **Family Academic Support Controls** do not play an economically significant role. **Gaming & Internet Use Controls** (# of video gaming systems at a student's home, and parental permission for playing video games or recreational internet use on weekdays) collectively play a small role in explaining leisure preferences. Adding gender-race interactions and gender-school-district interactions to specification 3 does not meaningfully change point estimates. **St. Dev. Effect** represents the change in standard deviation units of $\log(\theta_l)$ from switching the value of a binary regressor from 0 to 1 or from increasing the value of a continuous regressor by one standard deviation.

Similar patterns do not emerge for motivation level θ_l, with school district having no significant effect on time preferences beyond what is predicted by other factors such as gender, race, neighborhood socioeconomic traits, and a rich set of covariates on preferences, attitudes, consumption level, and outside options for time use. Finally, our Tobit results also speak to a classic question of whether better outcomes at higher-performing schools are due primarily to treatment by more advantageous school inputs or to selection of more academically adept students onto their rolls. We indeed find that higher-performing schools benefit from significant advantageous selection on both θ_e and θ_l (see Figure 19, Online Appendix A). Below we further investigate whether/how schools produce value added in the learning process.

There are several other insights that emerge from our decomposition of unobserved student characteristics. Reporting math as a favorite subject is unsurprisingly predictive of a significant increase in willingness to spend time on math, though it is also interesting, and perhaps reassuring, that listing math as one's least favorite subject is *not* a significant predictor of lack of motivation. We also find that being either more intrinsically motivated or more extrinsically motivated are *both* strong indicators of responsiveness to our extrinsic financial incentives for students to divert extra leisure time toward math

activity. 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, Hickman, & List, 2019), rather than a conflicting role as previously thought (e.g., Gneezy & Rustichini, 2000; Bénabou & Tirole, 2003; Leuven, Oosterbeek, & van der Klaauw, 2010).

We also assess the relationship between neighborhood socioeconomic traits and the current values of θ_e and θ_l. 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 resource deprivation. While affluence plays no meaningful role in determining θ_e and θ_l, resource deprivation is a statistically and economically significant predictor of a child being *less* motivated for academic pursuits.[30]

Figure 6 displays the selection-corrected distributions of θ_l and θ_e by gender for the entire sample population (regardless of worker status), using Tobit model estimates. The CDFs graphically depict the gender differences explained above; namely, that females tend to have lower academic efficiency but also lower time preference with regard to math activity, relative to males. Interestingly, in the case of the gender comparison, the motivation factor dominates in terms of total work volume on our website. While the average for males is 8.5 quizzes completed, females completed 35% more (11.5 quizzes), despite taking longer on each. This difference is significant at conventional levels ($p-value = 0.001$). A similar pattern emerges in survey data on daily homework times (all academic subjects) as well: females self-report 1.31 hours per day on homework activities, which constitutes a significant ($p-value = 0.0004$) increase of 10% relative to males, at 1.19 hours per day. In short, our descriptive and causal results all indicate that, conditional on environmental factors, attitudes, and preferences, while males seem to have a comparative advantage of academic efficiency in mathematics, females have a mathematics comparative advantage in terms of work ethic.

Figure 7 depicts the selection-corrected distributions of θ_e and θ_l by race. The distribution of $\log(\theta_e)$ is strikingly shifted to the right for Black and Hispanic students compared to Whites/Asians; the gap being several times larger than the analogous gender gap. The most motivated (i.e., lowest θ_l) Black and Hispanic students require fewer incentives to engage in extra math study, relative to the most motivated White/Asian students. Among the least motivated, Blacks and Whites/Asians look very similar, but the least motivated (i.e., highest θ_l) two-thirds of Hispanics lag significantly behind the other two groups in responsiveness to external incentives, for a given academic efficiency level. Two facts from Tables 1 and 2 provide a possible explanation for why: first, within our sample population Hispanics are most heavily represented in District 3; second, District 3 has the highest proportion of students with limited English proficiency. This is suggestive that linguistic barriers may play a significant role in reducing academic motivation for children from Hispanic immigrant families. An exploration of linguistic barriers is beyond the scope of this project, but it underscores an important consideration when interpreting the race parameters in Tables 4 and 5: these terms need not represent anything innate about a child due to his/her race, but may instead be a proxy for other cultural, social, or linguistic factors not captured by our model. All of these considerations are important questions deserving further attention in future research.[31]

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

[31] Bodoh-Creed and Hickman (2017) structurally estimate unobserved student traits using observational data on college admissions. In their data, race no longer retains predictive power for unobserved student characteristics, conditional on parents' income, wealth, education, and marital status.

FIGURE 6. Distributions of Characteristics by Gender

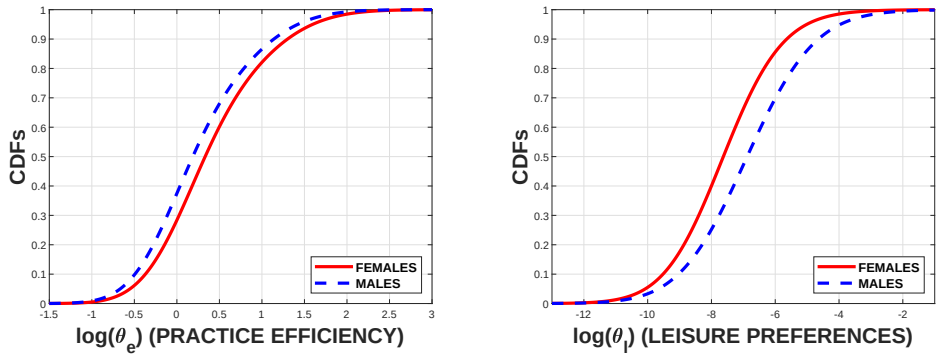
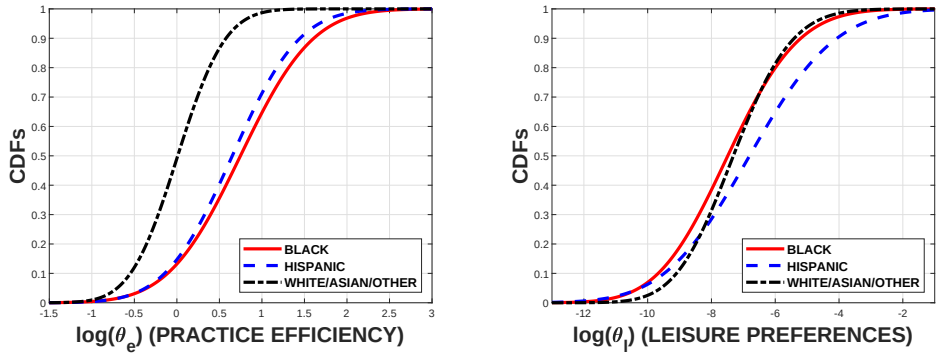


FIGURE 7. Distributions of Characteristics by Race



5.3. **Determinants of initial math skill.** In Sections 2.3 and 4.3, we formalize long-run development of math skill (measured by a standardized pre-test) as a Cobb-Douglass production process (equation (14)) where the principal inputs that schools use to produce new learning are student traits (θ_e,θ_l). The total factor productivity (TFP) term and the production shares of the two main inputs are idiosyncratic to each student i, depending on a vector of external factors $\boldsymbol{W}_i$ which include school quality, gender, race, family support controls, and socioeconomic proxies. Intuitively, holding a child's traits $(\theta_{ei}, \theta_{li})$ fixed, each of these external factors is allowed to play a direct role in the production process—through altering TFP A_i—as well as an indirect role—by altering the effectiveness of the primary inputs through the production shares α_{ei} and α_{li}.[32]

Empirical results are presented in Table 6. 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 6 provides several interesting insights. First, we find that both θ_e and θ_l are important determinants of initial math skill, but θ_e 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_l, on average. Together, these results suggest that educational interventions, such as Fryer (2011), Levitt et al. (2016),

[32]When interpreting empirical results, recall that θ_e and θ_l 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.

TABLE 6. COBB-DOUGLAS PRODUCTION OF INITIAL MATH PROFICIENCY

SPECIFICATION: **DEPENDENT VARIABLE:** $\log(S_1)$	**(1)** (Mean; St. Dev.)	**(2)** (Mean; St. Dev.)	**(3)** (Mean; St. Dev.)	**(4)** (Mean; St. Dev.)
TFP $(\widehat{\log(A_i)})$	(2.973; 0)	(2.910; 0.149)	(2.872; 0.207)	(2.871; 0.214)
θ_e **Prod. Share** $(\widehat{\alpha}_{ei})$	$(-0.453; 0)$	$(-0.327; 0.107)$	$(-0.283; 0.106)$	$(-0.283; 0.105)$
θ_l **Prod. Share** $(\widehat{\alpha}_{li})$	$(-0.043; 0)$	$(-0.037; 0.007)$	$(-0.040; 0.020)$	$(-0.040; 0.020)$
	Mean St. Dev. Effect	Mean St. Dev. Effect	Mean St. Dev. Effect	Mean St. Dev. Effect
$\log(TFP)$ *(joint p-value)*	N/A	0.3543*** $(<10^{-16})$	0.4935*** $(<10^{-16})$	0.5085*** $(<10^{-16})$
$\log(\theta_e)$ *(joint p-value)*	-0.6435*** $(<10^{-16})$	-0.4648*** $(<10^{-16})$	-0.4030*** $(<10^{-16})$	-0.4030*** $(<10^{-16})$
$\log(\theta_l)$ *(joint p-value)*	-0.1562*** (3.1×10^{-16})	-0.1338*** (4.2×10^{-15})	-0.1448*** $(<10^{-16})$	-0.1465*** (8.8×10^{-11})
CONTROL VARIABLES:				
District 2 $(\widehat{\alpha}_{01}, \widehat{\alpha}_{e1}, \widehat{\alpha}_{l1})$ *(joint p-value)*	—	-0.3922*** $(<10^{-16})$	-0.3063*** $(<10^{-16})$	-0.2952*** $(<10^{-16})$
District 3 $(\widehat{\alpha}_{02}, \widehat{\alpha}_{e2}, \widehat{\alpha}_{l2})$ *(joint p-value)*	—	-0.7704*** $(<10^{-16})$	-0.7526*** $(<10^{-16})$	-0.6618*** $(<10^{-16})$
Grade 5 $(\widehat{\alpha}_{03}, \widehat{\alpha}_{e3}, \widehat{\alpha}_{l3})$ *(joint p-value)*	—	—	-0.2267*** (4.8×10^{-11})	-0.2250*** (1.1×10^{-10})
Female $(\widehat{\alpha}_{04}, \widehat{\alpha}_{e4}, \widehat{\alpha}_{l4})$ *(joint p-value)*	—	—	-0.0383*** (8.5×10^{-6})	-0.0649*** *(0.0001)*
Black $(\widehat{\alpha}_{05}, \widehat{\alpha}_{e5}, \widehat{\alpha}_{l5})$ *(joint p-value)*	—	—	-0.2316*** *(0.0018)*	-0.2157*** *(0.0026)*
Hispanic $(\widehat{\alpha}_{06}, \widehat{\alpha}_{e6}, \widehat{\alpha}_{l6})$ *(joint p-value)*	—	—	-0.0555** *(0.0263)*	-0.0508* *(0.0576)*
log(Mean Nbhd Income) Controls	no	no	no	yes
Nbhd Uninsured Minor Rate Controls	no	no	no	yes
# Peer & Adult Helper Controls	no	no	no	yes
N	1,676	1,676	1,676	1,676
R^2	0.406	0.487	0.512	0.514
Adjusted R^2	0.405	0.485	0.506	0.506

Notes: **Mean St. Dev. Effect** is the total impact of a variable through both TFP (direct effect) and production shares of student inputs (interactions). For discrete variables **Mean St. Dev. 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., $(\widehat{\alpha}_{ek}, \widehat{\alpha}_{lk})$, k=1,. . . ,6) have the following joint p-values: 5.1×10^{-6} for **District 2**; 1.7×10^{-7} for **District 3**; 0.2412 for **Grade 5**; 0.0026 for **Female**; 0.1305 for **Black**; and 0.0343 for **Hispanic**.
The p-value for a joint exclusion of all neighborhood socioeconomic terms and helper terms is 0.6302.

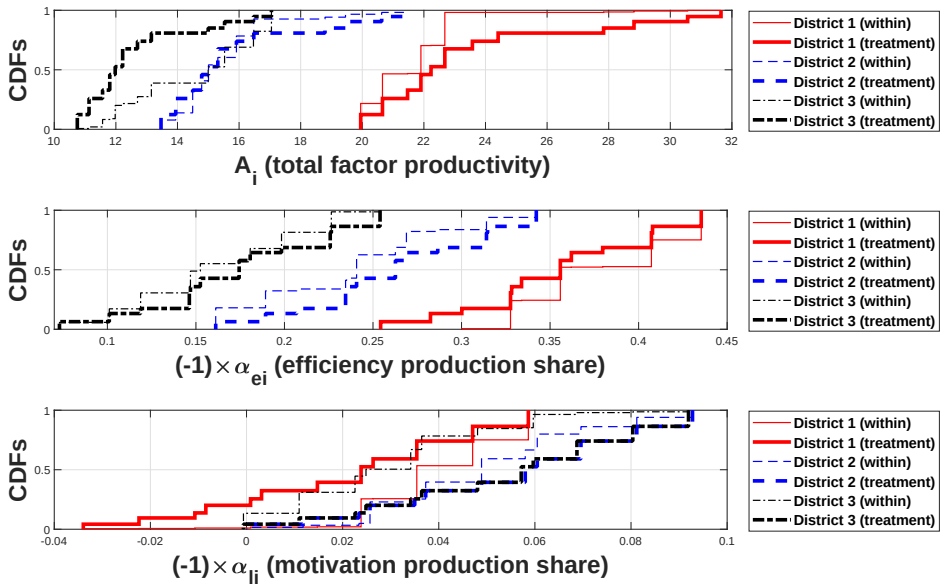
and Fryer et al. (2020), that aim to decrease gender or racial performance gaps in mathematics by motivating students through incentives or information about the returns to education may be misguided.[33] These groups already tend to be more motivated than their male or White/Asian peers, suggesting that motivation is not the primary barrier limiting their progress. Moreover, (in specification (4)) since TFP is 3.5 times as important as θ_l, and θ_e is 2.8 times as important, efforts to further incentivize marginal groups (further decreasing θ_l) will struggle to overcome the relative disadvantages these groups face.[34] We explore these considerations in more detail through counterfactual analyses in Section 6.

Second, we find strong evidence that school quality influences the production technology in important ways. The magnitudes of the school district effects again strongly conform to the pattern one might suspect from the suggestive evidence in Table 1: the difference between District 1 (the high performing

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

[34] 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, Nanowski, Nordstrom, 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.

FIGURE 8. Idiosyncratic Cobb-Douglas Production Parameters by School District



Notes: Since θ_e (θ_l) is inversely related to academic efficiency (motivation level), the associated production share α_e (α_l) is *negative*. The lower two panels multiply production shares α_e and α_l 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.

district) and District 2 (the middling school district) in terms of standard deviation effects is roughly half the difference between District 1 and District 3 (the struggling school district). Furthermore, the nature of the differences across school districts is not merely one of *levels*, but of the fundamental *shapes* of the production processes employed. Figure 8, 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 academic efficiency, 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.

Third, we also find evidence of decreasing returns to scale production technology in the sense that $-(\alpha_e+\alpha_l)$ 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.

5.4. **Determinants of incremental gains in math skill.** In Sections 2.4 and 4.3, we formalize improvements in math skill over the short run as a flexible quadratic polynomial (see equation (15)) in time spent on math activity (T) and volume of learning task completion (Q). Importantly, T and Q are also chosen by the student as functions of incentives and underlying characteristics (θ_e, θ_l), being micro-founded by the student choice model at the core of the field experimental design. The outcome variable of the short-run production function is the change in exam score ΔS between the post-exam and the pre-test (separated by 2-3 weeks of calendar time). Once again, we allow the intercept term and the slope coefficients on the primary productive inputs T and Q to be idiosyncratic, varying by the factors in $\boldsymbol{W}_i$, plus initial math skill S_1, academic efficiency θ_e, and time preference θ_l. This means that, in addition to the interactions from before, we are also allowing for student characteristics to play a dual role of

determining (T, Q) to begin with, and altering the rate at which learning task volume is converted into new math skill. The results of this analysis are presented in Table 7.

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.[35] Such is the case for T and Q (Table 3, Figure 2) where standard deviations exceed the respective 80^{th} 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 $pStDev_j \equiv F_j^{-1}(0.5|worker) - F_j^{-1}(0.159|worker)$, $j = t, q$, 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 Q (relative to all students, not just workers) are roughly 76 minutes of focused problem solving time and 8.4 website modules completed (i.e., 50.4 practice problems solved).

In Table 7 we find that completion of learning-by-doing tasks (and not simply time spent studying) is primarily responsible for short-term gains in mathematics proficiency. Notably, for inputs of time larger than the pStDev, T actually begins to play a negative role of tempering (but never swamping) the conversion rate of task completion into short-term gains in measured math proficiency. For example, the mean standard deviation effect of T, when computed relative to the usual standard deviation of time spent—at 154.5 minutes, being slightly more than double the pStDev—is -0.36 SD of ΔS.[36] These results are suggestive once again of a decreasing-returns-to-scale pattern in learning activity volume. We also see further evidence of a decreasing returns to scale production technology, but in a slightly different sense: the estimated standard deviation impact of pre-test score is significant (both economically and statistically) and negative. In words, 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. Note that the decreasing returns to scale insights from both long-run and short-run production technologies are also consistent with the remarkable degree of curvature that we find in the cost function: progress takes a lot of work (especially for high-θ_e types), and the increasing marginal costs of foregoing leisure time (due to θ_l and curvature in $c(\cdot)$) can very quickly become prohibitive.

We find that θ_e also alters the shape of the short-run learning technology in an economically meaningful way. That is, students with a more advantageous academic efficiency trait tend to not only accomplish more learning tasks per unit of time, but they also tend to derive more progress from those tasks in terms of measured math proficiency gains. 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 student covariates, school quality plays an important role in conversion of learning-by-doing activities into improvements in math proficiency over a short-run horizon. Moreover, the ordering among the three school districts is consistent with results from the previous two sections, though the difference between District 1 and District 2 is a bit smaller, relative to the District 1-District 3 comparison.

[35] 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.

[36] Importantly, one should keep in mind that all results in Table 7 are measured relative to extra-curricular math study over a fixed time window. Thus, the interpretation is that between 1.25 and 2.6 *extra hours* of math problem problem solving time *within a two-week window*, the role of time expenditure on learning progress switches from positive to negative.

TABLE 7. PRODUCTION OF INCREMENTAL GAINS IN MATH SKILL

SPECIFICATION: **DEPENDENT VARIABLE:** ΔS	**(1)** (Mean; St. Dev.)	**(2)** (Mean; St. Dev.)	**(3)** (Mean; St. Dev.)	**(4)** (Mean; St. Dev.)
Baseline 2-Week Gains w/$T_i = Q_i = 0$ $\left(\widehat{\Delta}_{0i}\right)$	$(0.495;\ 0)$	$(-0.217;\ 1.459)$	$(0.039;\ 1.903)$	$(0.0028;\ 1.933)$
	Mean St. Dev. Effect	Mean St. Dev. Effect	Mean St. Dev. Effect	Mean St. Dev. Effect
T **(standardized)**† $\left(\widehat{\Delta}_{1i}, \widehat{\Delta}_{2i}, \widehat{\Delta}_{5i}\right)$ *(joint p-value)*	-0.0013*** (5.5×10^{-5})	0.0187*** $(<10^{-16})$	0.0151*** $(<10^{-16})$	0.0165*** $(<10^{-16})$
Q† $\left(\widehat{\Delta}_{3i}, \widehat{\Delta}_{4i}, \widehat{\Delta}_{5i}\right)$ *(joint p-value)*	0.1950*** (5.4×10^{-8})	0.4385*** $(<10^{-16})$	0.5048*** $(<10^{-16})$	0.5378*** $(<10^{-16})$
S_1 **(standardized)** $(\widehat{\delta}_{0,1}, \ldots, \widehat{\delta}_{5,1})$ *(joint p-value)*	—	-0.4255*** $(<10^{-16})$	-0.4461*** $(<10^{-16})$	-0.4401*** $(<10^{-16})$
$\log(\theta_e)$ $(\widehat{\delta}_{0,2}, \ldots, \widehat{\delta}_{5,2})$ *(joint p-value)*	—	-0.2142*** (7.6×10^{-8})	-0.1776*** (3.4×10^{-10})	-0.1613*** *(0.0004)*
$\log(\theta_l)$ $(\widehat{\delta}_{0,3}, \ldots, \widehat{\delta}_{5,3})$ *(joint p-value)*	—	0.0284 *(0.1393)*	0.0505 *(0.3714)*	0.0601* *(0.0956)*
District 2 $(\widehat{\delta}_{0,4}, \ldots, \widehat{\delta}_{5,4})$ *(joint p-value)*	—	-0.2036*** *(0.0064)*	-0.2210*** (1.0×10^{-6})	-0.1286*** *(0.0026)*
District 3 $(\widehat{\delta}_{0,5}, \ldots, \widehat{\delta}_{5,5})$ *(joint p-value)*	—	-0.4733*** $(<10^{-16})$	-0.5340*** $(<10^{-16})$	-0.4271*** *(0.0094)*
Grade 5 $(\widehat{\delta}_{0,6}, \ldots, \widehat{\delta}_{5,6})$ *(joint p-value)*	—	—	-0.2401*** (6.5×10^{-7})	-0.2430*** (1.0×10^{-6})
Female $(\widehat{\delta}_{0,7}, \ldots, \widehat{\delta}_{5,7})$ *(joint p-value)*	—	—	0.0508* *(0.0530)*	0.1042** *(0.0362)*
Black $(\widehat{\delta}_{0,8}, \ldots, \widehat{\delta}_{5,8})$ *(joint p-value)*	—	—	0.0287*** (1.6×10^{-14})	0.0308*** (4.3×10^{-6})
Hispanic $(\widehat{\delta}_{0,9}, \ldots, \widehat{\delta}_{5,9})$ *(joint p-value)*	—	—	-0.0207* *(0.0696)*	0.0051** *(0.0152)*
log(Mean Nbhd Income) Controls	no	no	no	yes
Nbhd Uninsured Minor Rate Controls	no	no	no	yes
# Peer & Adult Helper Controls	no	no	no	yes
N	1,494	1,494	1,494	1,494
R^2	0.095	0.200	0.222	0.230
Adjusted R^2	0.092	0.181	0.190	0.188

Notes: **Mean St. Dev. 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 St. Dev. 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*: 9.7×10^{-7} for S_1 (standardized pre-test score); 0.0031 for $\log(\theta_e)$; 0.0812 for $\log(\theta_l)$; 0.0092 for **District 2**; 0.0587 for **District 3**; 0.0005 for **Grade 5**; 0.0458 for **Female**; 1.9×10^{-6} for **Black**; and 0.0135 for **Hispanic**.
Neighborhood socioeconomic proxies are statistically significant (*joint p-values* of *(0.0031)* and *(0.0437)*, respectively) but play a small role: a simultaneous one-standard-deviation improvement in both **log(Mean Nbhd Income)** and **Nbhd Uninsured Minor Rate** is predicted to result in only a 5.88% standard deviation increase in ΔS.
†Due to heavily skewed distributions of T and Q, 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 and standard deviation are the same.

In interpreting the results from Table 7 regarding standard deviation effects of T and Q, one should keep in mind that they involve many complicated interactions between variables. For example, the mean (across all students) predicted standard deviation effect of Q is roughly 2.7 exam score points (on a 40-point scale), or roughly 19 practice problems solved (with interactive feedback) per exam score point of improvement. However, for children at different school districts, with different initial proficiency, with different unobserved traits, and/or with different home background and demographic variables, the personalized prediction can vary somewhat. One encouraging aspect of model estimates for policymakers and education practitioners is that following pStDev=8.4 completed modules of extra math activity, the raw, pair-wise Kendall's rank correlations between the predicted shift ΔS and θ_e/θ_l are actually *positive* (0.4 and 0.3, respectively), and for pre-test score the rank correlation is *negative* (-0.31). In plain English, we learn an important lesson from this exercise: learning mathematics is accessible to anyone in the sense that there are enough other mitigating factors so that having a less advantageous latent characteristic θ_e or θ_l, or low initial math skill, need not bar any student from making progress.

FIGURE 9. Counterfactual Achievement Gaps: Black vs White/Asian

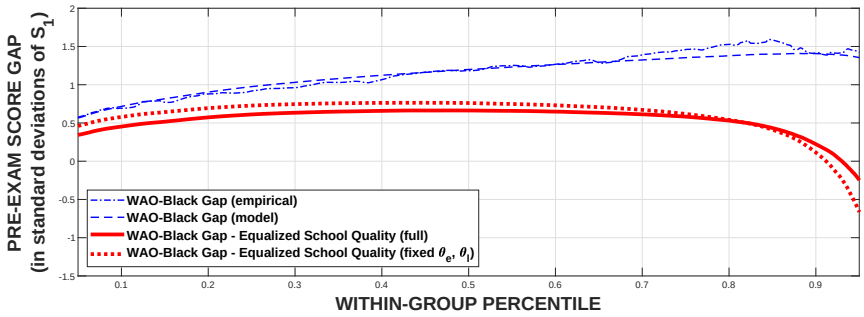
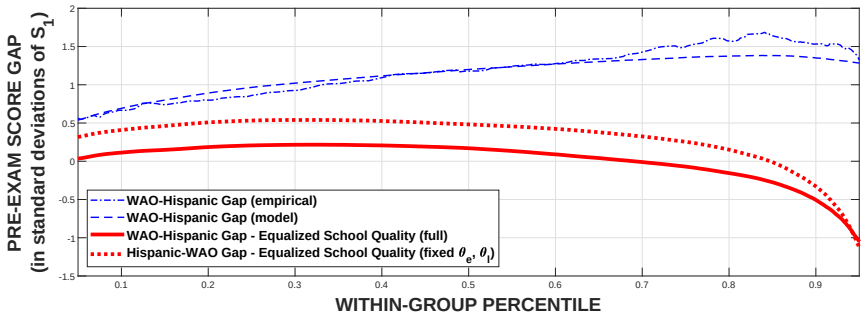


FIGURE 10. Counterfactual Achievement Gaps: Hispanic vs White/Asian



Notes: for $r \in (0.05, 0.95)$ Figure 9 (Figure 10) 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.

6. Counterfactual Analysis

In this penultimate section, we 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.[37] We then use model estimates to compute adjusted θ_e^* under the new school assignments, and we simulate counterfactual distributions of pre-exam scores and choices of T and Q 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.

6.1. **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 9 and 10, and numerically in Table 8. 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

[37] 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.

TABLE 8. SCHOOL-QUALITY EQUALIZATION: LONG-RUN ACHIEVEMENT GAPS

	PERCENT CHANGE IN ACHIEVEMENT GAPS AT:					
	10^{th} Percentile	25^{th} Percentile	Median	75^{th} Percentile	90^{th} Percentile	Mean Integrated %Chng.
Black *(full schl. qual. equalization)*	−36.8%	−37.3%	−44.7%	−57.0%	−84.2%	−45.0%
Black *(fixed (θ_e, θ_l))*	−19.2%	−25.4%	−36.6%	−54.3%	−91.9%	−38.9%
Hispanic Students *(full schl. qual. equalization)*	−83.6%	−78.6%	−85.8%	−105.2%	−137.4%	−85.8%
Hispanic Students *(fixed (θ_e, θ_l))*	−40.9%	−44.9%	−59.8%	−81.2%	−124.3%	−60.7%

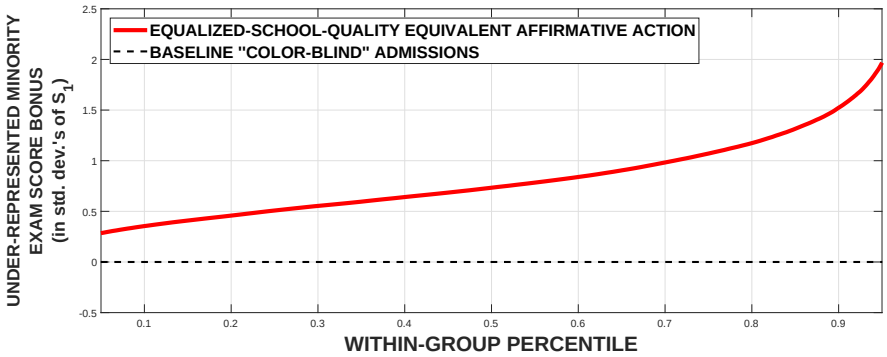
for Whites/Asians would cause the highest performing Black and Hispanic students to actually overtake their White/Asian counterparts in terms of exam score performance. Integrating over gap closure magnitudes at different percentiles generates a single aggregate summary value: holding all other student characteristics fixed, racial differences in school quality account for roughly 45% of the achievement gap between Blacks and Whites/Asians in our sample, and roughly 85% 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_e 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 (86% and 71% of the narrowing, respectively) is due to changes in the long-run production technology that exist at higher-quality schools, holding student traits fixed.

6.1.1. *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, Hickman, and Price (2020a, 2020b), 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 11. 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.[38] Note that the plot zooms in on the 5-95 range since behavior in

[38]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 11. 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.

the extreme tails for 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.

6.2. **Incentive response counterfactuals.** Finally, we seek to better understand the extent to which a policy-maker 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 Q (or T alternatively) under different contracts. For example, the White/Asian *Incentive Response Function* for a contract 1-to-contract 2 shift would be

$$IRF(j, W/A, 1, 2) \equiv F_j^{-1}(r|W/A, contract\ 2) - F_j^{-1}(r|W/A, contract\ 1),\ j = q, t,\ r \in [0, 1], \tag{16}$$

or the quantile function of Q 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.

FIGURE 12. Incentive Response Gaps in Learning Activities: Black vs White/Asian

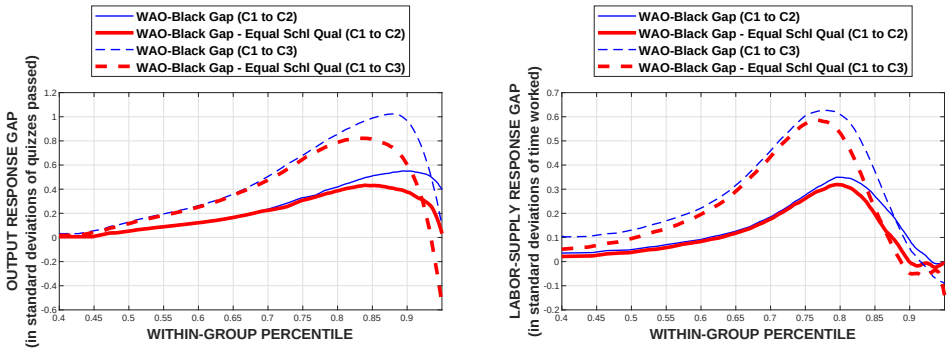
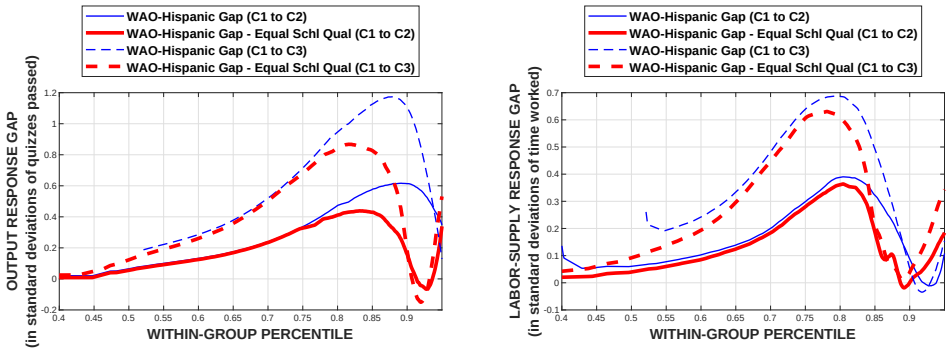


FIGURE 13. Incentive Response Gaps in Learning Activities: Hispanic vs White/Asian



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.

From our earlier analysis, one might believe that since Black students have systematically lower values of time preference θ_l, 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 time value *and* how much time is needed for task completion, which is a function of θ_e. While it is true that a lower θ_l makes it less burdensome for a student to give up an hour of would-be leisure time, higher values of θ_e 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_e 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 quiz output Q and the right panels show time worked T. Incentive responses and response gaps are fairly low until the 75^{th} percentile (i.e., most studious) students. In that upper region the response gaps in terms of Q are quite substantial, but are reduced significantly by equalizing school quality, with its implied increase of academic efficiency (i.e., reduction in θ_e). 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 $c(t; \widehat{\pi}_c)$, 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.

FIGURE 14. Incentive Response in Learning Activities: Black

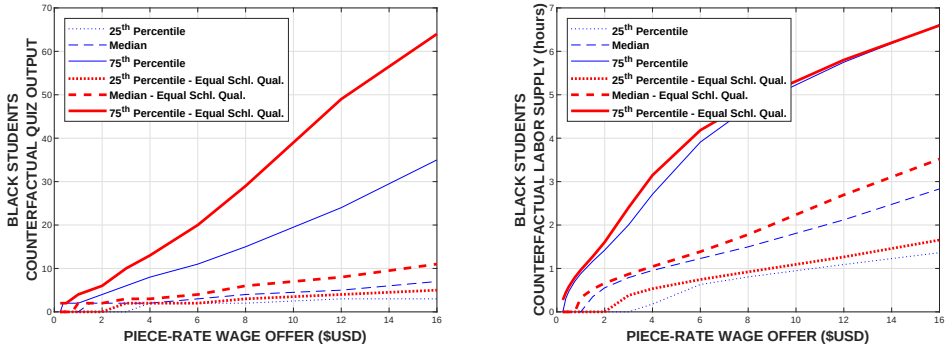
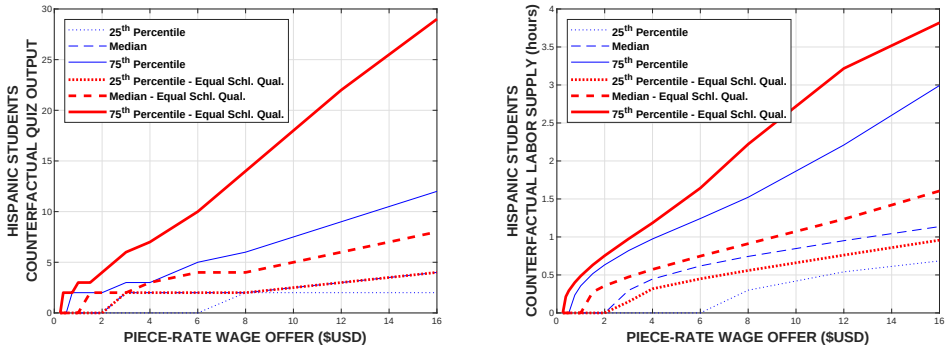


FIGURE 15. Incentive Response in Learning Activities: Hispanic



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 quiz 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_l *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_l. However, for both groups overcoming their disadvantage in terms of θ_e 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 academic efficiency θ_e while leaving θ_l 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.

7. Conclusion

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 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.

Our study contributes to the literature by providing insights into human capital formation and its determinants during one phase of the education process. Our approach is unique in that it uses a field experiment to identify key components of a structural model that illuminate the relationship between time and adolescent skill formation. By designing and operating our own web-based learning platform we are not only able to expose students to controlled variation in incentives, but we also gain a unique window into the temporal profile of study time supplied, volume of learning task completion, and how these inputs map into measured subject proficiency. In doing so, we discuss new interpretations of motivation, provide a novel view of policies that are geared toward opportunity versus achievement, and develop a contemporary view of optimal approaches to lessen racial and gender achievement gaps during the adolescent years (see Kautz et al., 2017; Joensen & Nielsen, 2016; Joensen et al., 2020, for other work on skill formation in the adolescent years).

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 their performance. We show that groups of students, whether defined by race, gender, or school district, who are under-performing in mathematics tend not to be any less motivated (and several are more motivated) compared to groups who on average perform better. Rather, these under-performing groups tend to have lower academic efficiency, meaning that even when they put in time studying, they struggle more than others to convert this time into academic success. Further increasing their motivation to put in time does not address this issue, as the amount of additional time that is required to close the performance gap is very costly to the student and likely infeasible to achieve. The effective closure of performance gaps between under-represented minority students and their counterparts, for example, cannot feasibly rely on efforts to better motivate students, but would rather need to address the differences in academic efficiency, which are driven by factors such as school quality and resource deprivation/poverty.

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 & 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.

Angrist, J., Lang, D., & Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, *1*, 136–163.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Becker, G. S. (1964). *Human capital*. Columbia University Press: New York, NY.

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis with special reference to education, 3rd ed.* Chicago: University of Chicago Press.

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

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

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.

Bettinger, E., & Slonim, R. (2007). Patience among children. *Journal of Public Economics*, *91(1)*, 343–363.

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

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*.

Bodoh-Creed, A., & Hickman, B. R. (2017). Pre-College Human Capital Investment and Affirmative Action: A Structural Policy Analysis of US College Admissions. *Becker-Friedman Institute Working Paper Series*, *No. 2017-11*.

Bodoh-Creed, A., & Hickman, B. R. (2018). College assignment as a large contest. *Journal of Economic Theory*, *175*, 88–126.

Bowles, S. (1970). Toward and educational production function. In W. L. Hansen (Ed.), *Education, income and human capital.* (pp. 11–60). National Bureau of Economic Research.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21–32.

Bruner, J. S. (1966). *Toward a theory of instruction*. New York: Norton.

Bruner, J. S. (1985). Models of the learner. *Educational Researcher*, *31*, 21–32.

Cappelen, A., List, J., Samek, A., & Tungodden, B. (2020). The effect of early-childhood education on social preferences. *Journal of Political Economy*, *128*(7), 2739–2758.

Carroll, J. B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp. 87–136). University of Pittsburgh Press.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *64*, 723–733.

Carroll, J. B. (1989). The carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, *18*, 26–31.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*, 2633–2679.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, *106*, 855–902.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2009). The academic achievement gap in grades 3 to 8. *Review of Economics and Statistics*, *91*, 398–419.

Coleman, J. S., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.*

Cordes, S. A., Schwartz, A. E., & Stiefel, L. (2019). The effect of residential mobility on student performance: Evidence from new york city. *American Educational Research Journal*, *56*, 1380–1411.

Cotton, C. S., Hickman, B. R., & Price, J. P. (2020a). Affirmative action and human capital investment: Evidence from a randomized fieldexperiment. *Journal of Labor Economics*. (forthcoming)

Cotton, C. S., Hickman, B. R., & Price, J. P. (2020b). Affirmative action, shifting competition, and human capital accumulation: A comparative static analysis of investment contests. *Queen's University working paper*.

Cotton, C. S., Nanowski, J., Nordstrom, A., & Richert, E. (2020). Improving girls education outcomes through community-wide information and empowerment campaigns. *Queen's University Working Paper*.

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.

Currie, J. (2009). Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and

human capital development. *Journal of Economic Literature*, *47(1)*, 87–122.

Denham, C., & Leiberman, A. (Eds.). (1980). Washington, DC: National Institute of Education.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, *3*, 357–370.

D'Haultfoeuille, X., & Février, P. (2015). Identification of triangular nonseparable models with discrete instruments. *Econometrica*, *83(3)*, 1199–1210.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087–1101.

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. G., Jr. (2011). Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics*, *126*, 1755–1798.

Fryer, R. G., Jr., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*, 447–464.

Fryer, R. G., Jr., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*, 210–240.

Fryer, R. G., Jr., Levitt, S. D., & List, J. A. (2015). Parental incentives and early childhood achievement: A field experiment in chicago heights. *NBER Working Paper No. 21477*.

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.

Goldin, C. (2016). Human capital. In C. Diebolt & M. Haupert (Eds.), *Handbook of cliometrics.* (pp. 55–86). Springer Verlag: Heidelberg, Germany.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, *320*, 1164–1165.

Hanushek, E. A. (2020). Education production functions. In S. Bradley & C. Green (Eds.), *The economics of education (second edition).* (pp. 161–170). Academic Press.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Disruption versus tiebout improvement: The costs and benefits of switching schools. *Journal of Public Economics*, *88*, 1721–1746.

Hanushek, E. A., & Rivkin, S. G. (2006). School quality and the black-white achievement gap. *NBER Working Paper no 12651*.

Hanushek, E. A., & Rivkin, S. G. (2009). Harming the best: How schools affect the black-white achievement gap. *Journal of Policy Analysis and Management*, *28*, 366–393.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47(1)*, 153–161.

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic Inquiry*, *46*, 289–324.

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*.

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

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139–155.

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*, 8801–8807.

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.

James, W. (1890). New York: Henry Holt.

Joensen, J. S., List, J. A., Samek, A., & Uchida, H. (2020). Using a field experiment to understand skill formation in adolescence. *working paper*.

Joensen, J. S., & Nielsen, H. S. (2016). Mathematics and gender: Heterogeneity in causes and consequences. *Economic Journal*, *126*, 1129–1163.

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.

Katz, L. F., Kling, J. R., & Liebman, J. B. (2001). Moving to opportunity in boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics*, *116*, 607–654.

Katzman, M. T. (1967). Distribution and production in a big city elementary school system. *Dissertation, Yale University*.

Kautz, T., Heckman, J. J., Diris, R., ter Weel, B., & Borghans, L. (2017). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. *NBER Working Paper no 20749*.

Kiesling, H. J. (1968). *High school size and cost factors*. U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research.

Kosse, F., Deckers, T., Pinger, P., Schildberg-Hörisch, H., & Falk, A. (2020). The formation of prosociality: causal evidence on the role of social environment. *Journal of Political Economy*, *128*(2), 434–467.

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 17 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.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, *66(4)*, 281–302.

Morrison, H. C. (1926). *The practice of teaching in the secondary school*. The University of Chicago Press.

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

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

OECD. (2015). *The ABC of gender equality in education*. (Technical Report)

Rumberger, R. W. (2015). Student mobility: Causes, consequences, and solutions. *National Education Policy Center Policy Brief*.

Schultz, T. (1961). Investment in human capital. *American Economic Review*, *51(1)*, 1–17.

Schunk, D. H. (2020). Pearson.

Schwartz, A. E., Horn, K. M., Ellen, I. G., & Cordes, S. A. (2020). The effect of residential mobility on student performance: Evidence from new york city. *Journal of Policy Analysis and Management*, *39*, 131–158.

Schwartz, A. E., Stiefel, L., & Cordes, S. A. (2017). Moving matters: The causal effect of moving schools on student performance. *Education Finance and Policy*, *12*, 419–446.

Seo, H. K. (2020). Disinterested, lost, or both? Estimating productivity thresholds using and education field experiment. *Univesity of Chicago Working Paper*.

Skinner, B. F. (1954). The science of learning and teh art of teaching. *Harvard Educational Review*, *24*, 86–97.

Su, C.-L., & Judd, K. L. (2012). Constrained optimization approaches to estimation of structural models. *Econometrica*, *80(5)*, 2213–2230.

Titchener, E. B. (1909). New York: Macmillan.

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

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wang, M.-T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancyâĂŞ-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.

Appendix A. ONLINE SUPPLEMENT: Additional Tables and Figures

Table 9. 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.

Table 10. DEMOGRAPHICS BY CENSUS BLOCK GROUP

Variable	EXPERIMENTAL SAMPLE	ILLINOIS STATE
Mean Nbhd Hshld Income:		
weighted mean	$101,698	$71,602
weighted 5-95 range	[$35K,$156K]	[$30K,$128K]
Mean Nbhd Home Value:		
weighted mean	$361,935	$198,786
weighted 5-95 range	[$94K,$723K]	[$69K,$432K]
HS Graduation Rate (Adults 25+):		
weighted mean	0.9149	0.857
weighted 5-95 range	[0.58,1]	[.57,0.99]
Col Grad Rate (Adults 25+):		
weighted mean	0.5364	0.294
weighted 5-95 range	[0.05,0.92]	[0.04,0.72]

Notes: There are 9691 block groups in the state of Illinois. Our study sample consists of 161 census block groups in total. All variables described in this table are measured at the neighborhood (Census block group) level. Means are weighted by headcount of students residing in each Census block group. 5-95 range is weighted by headcount of students residing in each Census block group.

FIGURE 16. Conditionally Heteroskedastic Work Time Shocks

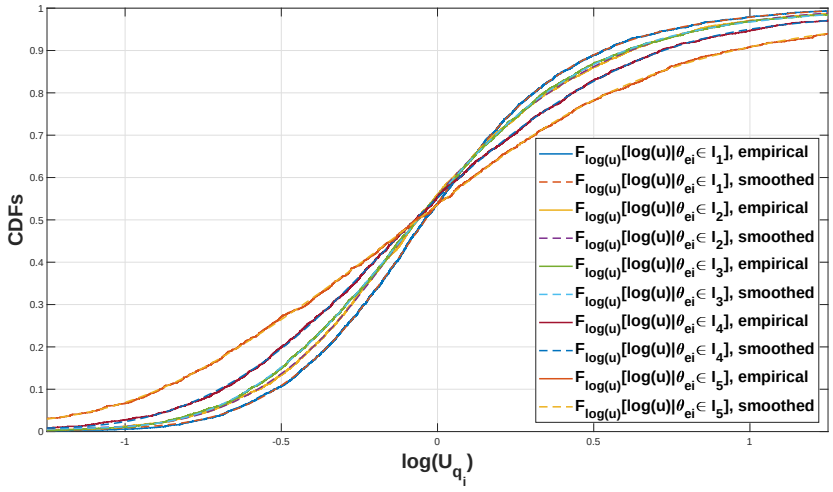


FIGURE 17. Upper Tail Extrapolation

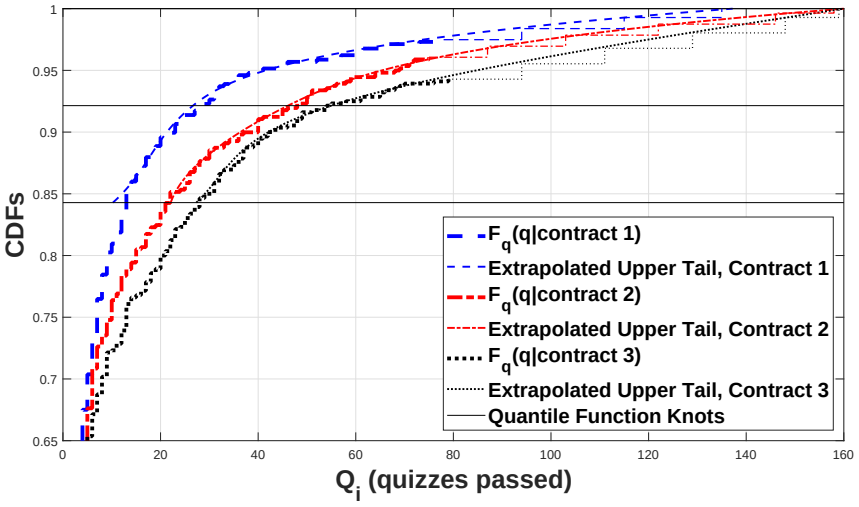


FIGURE 18. Cost Model Fit

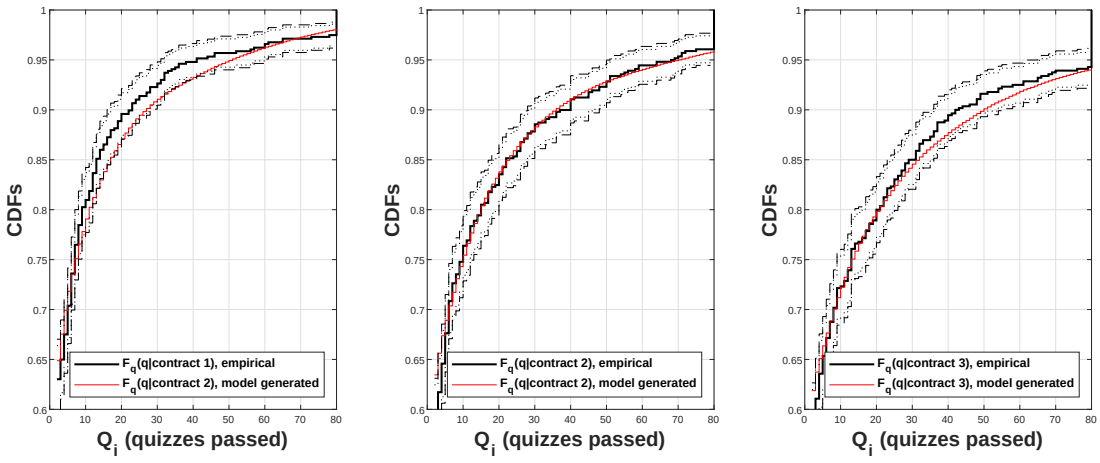
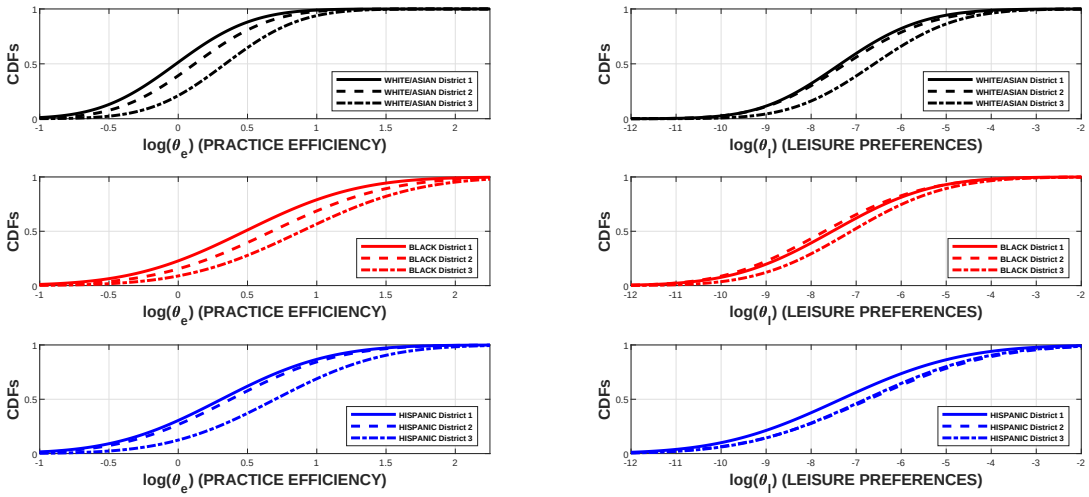


FIGURE 19


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